ITR: Detection and Localization of Anomalous Network Behavior

A. Hero(PI), P. Amaranth(co-PI), R, Baraniuk(co-PI), A. Benveniste(Collaborator)
M. Coates(Collaborator), R. Dwarshuis(Collaborator), P. Honeyman(co-PI),
S. Lafortune(co-PI), M. Liu(co-PI),
G. Michailidis(co-PI), B. Noble(co-PI), R. Nowak(co-PI),
J. Ogden(co-PI) S. Pradhan(co-PI), A. Prakash(co-PI),
R. Riedi(co-PI), D. Teneketzis(co-PI), D. Wallach(co-PI)

March 29, 2002

1 Summary

Today's private and public communications networks are critical systems of data terminals, routers, and switches which 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 and pattern recognition on a global 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 distributed pattern recognition approaches for detecting and identifying pattern changes; 3) application to automated distributed detection of intrusions, denial of service (DDoS) attacks, 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 flow statistics throughout the network and to characterize what constitutes a significant change in flow patterns. However, the high-dimensional 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 decentralized detection. With the help of our industry 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 data obtained from data collection sites at Internet2 and Merit Network, Inc., we will collect complete data traces from switches and hubs using a PacketVault technology 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.

Our approach to anomaly detection and localization is a potent combination of emerging techniques in pattern recognition, decentralized detection, and discrete event systems. We will develop new classes of flexible hybrid pattern recognition algorithms to correlate events over space-time that are scalable to large volumes of data acquired from global Internet measurements. While our focus will be on off-line implementations, we will explore implementations that can be used to generate and correlate alerts in real-time with a minimum of human intervention. Our global approach goes well beyond previously introduced techniques of fault detection, traffic analysis, and alert correlation which have been restricted to much smaller scale problems.

This project will involve precollege, college, and continuing education. A new cross-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 lab projects. We will also sponsor a summer internship program in networking for 20-30 qualified high school and middle school students. These students will participate in various educational and recreational signal processing and networking activities which we will organize. An educational innovation of our project is the development of an isolated networked environment at Rice and Michigan for emulation of attacks. A yearly summer competition will pit Rice students against Michigan students in a contest for most effective detection and response to the simulated attack. This laboratory will serve dual purposes: 1) an exciting environment for learning about network security; and 2) a controlled testbed to explore different algorithms developed in our research. Our industry collaborators will provide guidance and inputs to this project including: help in emulating realistic attack scenarios, evaluating response strategies, guidance on large-scale testing of our pattern recognition algorithms, and deployment of data collection sites.

1. Executive Summary

The principal aim of this project is to study, develop and disseminate methodologies for distributed data collection and aggregation to be used for rapid detection 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 two Internet service providers in the relevant fields of pattern recognition, decentralized detection and control, multivariate statistics, network failure detection and diagnosis, network security, and data collection. The project proposes four inter-related areas of activity: distributed data collection; large-scale pattern recognition; networking applications; and network security education. These activities are discussed below.

Distributed data collection: Accurate detection of spatially distributed flow-pattern changes requires access to data collection devices at many different sites in the network. We will deploy a novel combination of data collection devices on the Internet. Aggregated data will be collected from routers on MichNet and other Internet2 networks, controlled by our collaborators at Merit Network (Inc) and Internet2 (Inc). In addition to this aggregated data, complete header flow 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 (a few dozen) 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 much of this data publically available in an easily accessible format through our website. We also intend to freely disseminate the end-to-end measurement software to any sites interested in participating in data collection for the project.

Distributed pattern recognition: Recognition of subtle spatially distributed patterns from a large number of collection sites is an extremely challenging problem. We launch a frontal attack on this problem which 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 take an approach incorporating elements of model-based and learning-based pattern recognition and anomaly detection. Particular innovations of our research are: 1) development of hybrid model-based and learning-based pattern recognition schemes that can be implemented in a decentralized manner; 2) application of discrete-event dynamic systems (DEDS) framework to emulate a baseline of events in the network; 3) integration of DEDS and non-linear time series models for learning (and detecting deviations from) the baseline from diverse types of network measurements; 4) development of particle filtering methods for data aggregation, detection, and feature classification.

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 the attacker attempst to flood the buffers of several servers on the network. We will cnsider detection of coorrdinated multiple-site intrusions involving robot larceny, i.e., theft of data and other files. Finally we will investigate service monitoring and verification applications using a combination of active probing and passive measurements.

Network Security Education: Improving computer security education at all levels (K-12, college, continuing ed) is a major goal of this project. Two peer-reviewed workshops will be organized on the topic of distributed data collection and network security. We will implement an undergraduate and graduate curriculum at Rice and UM which goes beyond the standard treatment (encryption, ...) and provides an interdisicplinary 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 at UM, Rice and elsewhere will match wits against each other on mitigating simulated attacks on their networks. The project also aims at improving K-12 education on computer security in general, and network security in particular. In collaboration with UM's camp CAEN, we will tailor the aforementioned computer game environment to high school kids enrolled in the two week summer camp. Finally, in collaboration with Merit Network, Inc., we will sponsor web-lectures, web-courses, and workshops for educating K-12 teachers and their technical support staff on computer security.

Several commercial and not-for-profit companies (Arbor Networks, Sprint, Lucent, Merit Network, Internet2, and Stanford Linear Accelerator (SLAC)) are interested in participating with us in this project - letters are attached. If funded, this project will likely result in major advances: 1) a fuller understanding of the limitations of global network inference methods for detecting and localizing potentially debilitating attacks and link failures; 2) an integrated and flexible multiple time series analysis methodololgy which can be employed in a decentralized manner; 3) scalable decentralized algorithms for detecting emerging attack patterns from routers switches and other data-collection sites; 4) dissemination of collected data software tool for distributed end-to-end data-collection; 5) instruction of undergraduates on computer security through a fun computer game which will serve to both generate data for validating our models and generating new patterns of attack and mitigation. An added benefit will be multi-disciplinary and practical training of graduate and undergraduate students in Signal Processing, Networking, Statistics, Discrete Event Systems, Optimization, and Security.

A diagram summarizing how the proposed effort is compartimentalized is given in Fig. 1. Principal associated co-PI's are listed with each of the activities,

2. Prior NSF Support

1. "Multiscale Signal and Image Processing using Singularity Grammars," NSF/CCR-9973188 (08/01/99 - 07/31/02)PI: Richard Baraniuk, Rice University

   Summary: This project aims to develop a framework for multiscale signal modeling, processing, and analysis that is matched to the 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 nonGaussian 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. Our multifractal traffic models are in use at a number of research laboratories and universities.

Education
- Global network security curriculum
- Networking security internships
- Networking laboratory
- Webcast seminars and workshops
- Summer UG and high school program
- Network gaming competition

(*Liu, Prakash, Wallach, Noble, Baraniuk, Hero*)

Data Collection
- Development and dissemination of software
- Multi-site information aggregation
- Distributed computation/compression/complexity
- Data collector authentication and privacy
- PacketVault usage

(*Prakash, Liu, Nobel, Wallach, Pradhan, Honeyman*
(Ogden, Dwarshuis, Amaranth)

Distributed Pattern Recognition and Detection
- State extraction from local data collectors
- State aggregation using discrete-event systems
- Spatio-temporal dynamical traffic models
- Response, false alarms, and QoS assurance
- Centralized vs. decentralized methods
- Performance mapping (channel id, tomography)

(*Teneketzis, Pradhan, Coates, Lafortune*
Liu, Nowak, Michailidis)

Applications
- Distributed denial of service (DDoS) attacks
- Multiple-site intrusions
- Robot larceny
- Service monitoring/verification

(*Lafortune, Reidi, Wallach, Nobel, Coates, Nowak, Hero*)

Figure 1: Major poles of research and education activities.

2. "Information theoretic analysis of tomographic systems," NSF/BCS-9024370(1993-1995), PI A.O. Hero, University of Michigan

   Summary: In this grant the fundamental limitations on performance of tomographic systems were characterized in terms of edge sensor placement, properties of the medium, and statistical variability of the measurements. This resulted in more pertinent criteria for design of tomographic data collection systems and in new high performance algorithms for reconstruction [29, 30, 27, 23, 54, 52, 51, 53, 28, 25], new iterative reconstruction algorithms [21, 20, 19] and performance analysis [26, 24, 22]. The paper [29] 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, (9/1/2000 to 8/31/2003), $205,000, S. Lafortune (PI) and D. Teneketzis (Co-PI).

   Summary: 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 where intermittent failures may occur. Our approach includes two major steps: building a discrete-event model of the system to be diagnosed, followed by construction of a *diagnostic protocol*, i.e., the set of communication and diagnostic decision rules employed by the various system diagnostic components for failure detection and identification. Our current research includes: (i) diagnosis of intermittent failures in the context of centralized architectures; (ii) dealing with communication delays in the context of coordinated decentralized architectures; (iii) development of protocols for failure diagnosis of distributed systems based on modular system models.

4. "Information Visualization through Graph Drawing: Modeling, Analysis and Optimization Is-

sues," NSF/IIS-9988095; 01/01/01-12/31/03, PI, George Michailidis, University of Michigan

Summary: 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. "A Framework and Methodology for Edge-Based Traffic Processing and Service Inference," ANI-0099148, (8/15/01 - 7/31/04), $966,704, 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. This project blends 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. The INCITE Project is developing new algorithms and implementations, providing a vital step towards better managing and understanding of Internet performance.

3. General Research Approach

The research approach described in this proposal represents a dramatic departure from existing activities in networking security, measurement, and mapping. Instead of focusing on parameter estimation, model fitting, or heuristic rule-based security measures, here we propose a pattern-theoretic approach to detection, localization, and classification of abnormal network behavior. The emphasis here is on the analysis of global patterns of network behavior, rather than local characteristics measured at a point. However, we are not suggesting that a global mapping or state of the network is necessary (or estimable). This project focuses on analyzing and detecting patterns lying in high dimensional feature spaces (collections of measurable network attributes) recorded at a large number of distributed sites.

The block diagram in Figure 2 provides a simplified illustration of our general distributed data-collection and pattern recognition framework with centralized data processing. Three local data collection sites (bottom of figure) tranmsit compressed (encoded) versions of features extracted from local data streams. The design of compression and the feature extraction algorithms can be based on models (multi-fractal time series), rules (learning) or grammers (DEDS). A central collection site (top of figure) 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 classifed 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).

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 aggregating 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

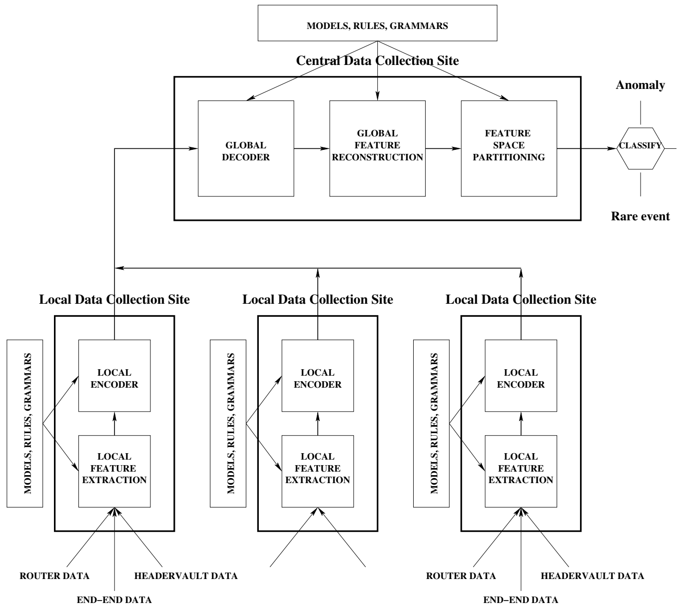

Figure 2: *Block diagram of our hierarchical decisionmaking approach for a single central collection site (top of diagram). There could in general be several layers intermediate collection sites which would successively aggregate data-collected at lower level sites. The decisionmaking is decentralized 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 decisionmaking is advantageous when a central site is not available or may itself come under attack.*

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. Pattern Recognition and Detection (Leader: Teneketzis)

We consider a hierachical network architecture consisting of two or more levels. 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 (the highest level of the hierarchy) or to intermediate-level nodes which are responsible for monitoring the operation of larger than local portions of the network. The intermediate nodes process the information they receive from local nodes and report to the network manager which is responsible for detecting behavior in the (whole) network or attacks on the networks.

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

4.1 Processing at local nodes

As relevant information is carried in the temporal variations of local traffic and packet flows, a stochastic dynamic system framework is natural for the local data collection sites. Such a framwork can yield a compact dynamical systems approximation to the "baseline" of the microscopic evolution of traffic and packet-level information (http/ftp requests, ping, netflow) at local nodes or small groups of neighboring nodes of the network. Application of stochastic dynamic systems framework could be model-based or learning-based. Model-based examples include variants on autoregressive moving average (ARMA) time series models such as: multifractal (MF) ARMA and fractional autoregresive integrated moving average (FARIMA) systems. Learning-based examples include non-parametric phase-space reconstruction algorithms using Taken's imbedding methods, classification and regression trees (CART), radial basis functions (RBF), or hierarchical clustering. 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. Measured features will be encoded to a small number of bits and tranmsitted to the higher levels of the network hierarchy. The encoding operation could simply classify the local measurements into one of several regimes of oepration, e.g. "baseline" or "anomaly," or it could encode to a more refined codebook using distributed vector quantization (VQ) techniques. The protocol of operation of each node as well as the representation of the dynamic systems describing the operation of individual nodes will be data-driven and will be continuously updated. The date-driven updates will be made so as to provide better quality of information to the higher levels of the network hierarchy.

4.2 Processing at higher levels of the network hierarchy

Upper levels of the data-colelction heirarchy will aggregate the received locally encoded features into global features as illustrated in Figure 2. Again, we will investigate both learning-based and model-based approaches to feature aggregation. In the model-based approach we will use a discrete event dynamic systems (DEDS) framework to capture the operation of the intermediate levels as well as the highest level of the aforementioned hierarchy.

The DEDS framework provides a low dimension description of what one might call the "macroscopic" evolution and operation of the network using formal language theory equipped with a grammar governing sequences of baseline events. As such, 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 of the hierarchy and the network manager receive information in the form of messages from lower levels of the hierarchy. Such messages may be ”events” (such as ”there is an increase in traffic in a certain part of the network”) or some ”statistic' of the information of lower hierarchical levels (such as the 'likelihood” of a traffic anomaly in a certain part of the network). As pointed out in Section 3.1.2, these messages describe deviations from the ”baseline' operation of lower levels of the hierarchy. Such deviations are caused by events that are ”unobservable” by the network. Examples of unobservable events include ”imitation of attacks on the network” or ”rare increases in normal traffic”. Based on the received messages, the intermediate levels of the hierarchy and the network manager have to estimate their ”aggregate” state and detect/diagnose attacks on the network. Estimation of the aggregate state is necessitated by the fact that the information about the network available to the intermediate levels of the hierarchy and the network manager is imperfect. 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. Detection of attacks will be based on the results of estimation.

The above framework can be represented by logical DEDS [7] or by Stochastic DEDS such as Markov or semi-Markov Chains [5]. The representation of these DEDS will be data-driven. The set of ”aggregate” states of the DEDS as well as the set of messages received by the network manager or the intermediate levels of the hierarchy will be continuosly updated as additional on-line information becomes available. The updates will be made so as to increase the likelihood of detection of attacks on the network.

4.3 Relationship between the approach to local nodes and the approach to higher hierarchical levels.

Our approaches to dealing with the operation of the various levels of the network hierarchy are tightly coupled, because decisions at any particular level affect the operation and decisions at other levels, as evidenced by the discussion below.

To improve the quality of its estimates (therefore, the quality of its diagnostic decisions) the network manager may querry lower levels of the hierarchy so as to acquire specific information. To provide the information requested by the network manager, the lower levels of the hierarchy may adjust their protocols (e.g. the rules of acquisition and processing of information; the classification of deviations from the baseline operation, hence, the quantization of their information). These protocol changes together with information received on-line lead to the adaptation of the dynamic systems (e.g. FARIMA, ARMA) describing the operation of local nodes. The information received by the network manager as a result of its queries to the lower levels of the network hierarchy may lead to adaptation of the DEDS describing the operation of the higher levels of the hierarchy. Thus, the two-way interaction/feedback 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 research issues at each level of the network hierarchy that are important/crucial to the overall network performance (that is measured by its ability to detect/diagnose attacks). We discuss these issues in the following sections.

4.4: Learning-based Approaches to Network Anomaly Detection

Learning approaches to pattern recognition have led to breakthroughs in high dimensional classification problems such as hand written character recognition, genetic sequencing, and image indexing [6, 14, **?**]. As examples, tree-based classifiers and kernel-based methods (Support Vector Machines - SVMs) are powerful and computationally efficient nonparametric classifiers. These classifiers do not require specification of a model; using sophisticated complexity-based regularization techniques, trees and SVMs are able to classify complex patterns in high dimensional spaces by learning from training examples alone.

Tree classifiers partition the feature space in a hierarchical manner, divide-and-conquer strategy, that enables robust and flexible pattern recognition. Tree classifiers have many desirable characteristics; they can easily handle "mixed" data types and missing data, they are robust to outliers and are insensitive to monotone transformations of the input features, and they are computationally scalable. Recently, we and others have devised a new approach to constructing tree classifiers that provides concrete bounds on the classification [**?**, **?**] performance (similar bounds are not available for the more well-known CART algorithm [**?**]). We have also applied CART like methods to universal prediction and reconstruction of non-linear time series [39] which can be easily adpated to constructing a baseline at the local data collection sites.

SVMs are another very powerful approach to learning-based pattern recognition [**?**]. Experimental results across a wide variety of high-dimensional pattern recognition tasks suggests that SVMs perform extremely well in many real world problems. The power of SVMs is that they convert non-linearly separable patterns to linearly separable patterns in a higher dimensional feature space where hyperplane classifiers can be applied to classify the patterns. 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 (with similar underlying distributional characteristics). The use of the VC complexity measure is also central to the new tree classifiers mentioned above.

To the best of our knowledge tree classifiers and SVMs 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?

- 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 dependencies/correlations 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 which collectively exploits the strengths of learning and modeling 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 anamolies, perhaps a nested sequence of hypotheses is a more robust approach to Internet pattern recognition. For example, anamolies could first be coarsely catergorized 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.)

4.5: Model-based Approaches

When they are available, one can always benefit from accurate physical models of the measurement process. Such models can specify a parametric marginal distribution for traffic flow or properties such as piecewise stationarity, independence, Markovianity, and other statistical attributes. Likewise, when attempting to detect or discriminate particular types of event sequences, e.g. anomalous combinations of sequences of packets types and sizes, formal grammers of events can provide useful models for the state of the network. In such cases discrete event systems (DEDS) models can be used to greatly reduce the unmanagable large dimension of the space of all admissible event sequences. In either case a good model can facilitate detection if it is scalable to the global network measurement scenario.

DEDS Approach

We propose to use logical and stochastic DEDS models that will work in tandem with the multivariate spatio-temporal traffic models, and Bayes importance sampling methods, described below, for the detection of component failures, attacks, and other anomalies in the network. The DEDS models will capture the network traffic behavior at a higher level of abstraction and on a different "macroscopic" time scale than the local traffic flow models. The events that will drive the DEDS models will comprise "observable" events that will be obtained by aggregating, quantizing, and filtering key variables from the spatio-temporal models as well as "unobservable" events that will capture special changes of the state of the overall network, possibly due to anomalies or attacks, that are not directly measured nor captured by the spatio-temporal models. The DEDS models, together with the sequences of observable events that will drive them online, will then be used to infer about the occurrence of the unobservable events and thus detect, if any, various anomalies in the behavior of the network.

Detection and Identification of Attacks Using DEDS Models

The theoretical foundations for the task of discrete-event model-based inferencing lie in the failure diagnosis methodology for logical DEDS developed in our prior investigations [45, 46, 44, 18, 36]. In fact, these works have been a major source of inspiration for related approaches for network alarm correlation and fault detection and isolation that have been successfully demonstrated in practical applications: large-scale telecommunication networks [1, 41, 3] and wireless LANs in vehicle platooning [16].

The methodology in [45, 18] will have to be significantly enhanced in order to address the objectives of this proposal. These enhancements include: (i) incorporation of nondeterministic and stochastic features in order to appropriately couple, in a hierarchical manner, the DEDS models with the data-driven spatio-temporal models; (ii) development of modular and distributed algorithmic implementations in order to address the scalability requirement; and (iii) incorporation of event-driven models of DDoS and spoofing attacks and development of distributed and asynchronous algorithms for the detection of such attacks.

According to the taxonomy proposed in [2], 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 [47, 55, 34] while signature detection assumes a partial model of "intrusion" [31, 35, 38, 8]. As its name indicates, compound detection assumes partial models of intrusive and normal behaviors [37]. This is the approach that we shall adopt in our investigations on model-based intrusion detection, consistent with our prior work in [45, 18].

In order to tackle the problem of intrusion detection by DEDS-model-based inferecing, we propose to retain the general features of the "DIAGNOSER" approach introduced in [45] but to generalize the notion of "failure-type labels" for tracking unobservable failure events to larger sets of labels that capture partial/complete attack patterns from a database of such patterns. In this manner, model-based inferencing will still be centered around the use of *diagnoser automata*, a highly desirable feature as it will allow for formal analyses of the diagnosability properties of classes of intrusions based on the available measurements (i.e., observable events). The construction and update of the generalized labeling function employed by a diagnoser automaton will be based on separate individual template automata for the various classes of attack patterns. This approach will have the advantage of modularizing the inferencing process and will permit reconfiguration as new attack patterns are discovered (leading to new template automata that will feed the diagnoser automaton). This will also allow our investigations to leverage on prior work based on the signature detection approach.

Local Traffic Flow Models

Given a particular configuration of the network, the flow of packets through the network generates traffic which might be characterized by a network of coupled multivariate time series models. These spatio-temporal traffic models capture the fast dynamics of local traffic flows as related to flows of traffic elsewhere in the network and exogenous inputs due to probing and ambient traffic sources. The parameters of the traffic model change according to the slower state transitions of the network which are determined by the DEDS model discussed above. We will investigate several classes of multivariate time series models including multiscale and multifractal traffic models [**?**] and [43] fractional autoregressive integrated moving average (FARIMA) models. There is evidence that these models can capture both the long term dependency as well as short range dependency of single stream IP traffic flows [32, 4, 40].

Multifractal models parsimoniously capture multiscale traffic features including self-similarity and long-range-dependence. These models are flexible, incorporating short- and long-term traffic correlations as well as high and low order moments. Parameters are easily computed through the efficient wavelet transform. In addition, the models are amenable to queuing analysis thus making them attractive for numerous applications. Several physical justifications for the multifractal models have been proposed including repartioning of traffic due to the protocol hierarchy and traffic multiplexing at routers.

However, several factors might play in favor of the FARIMA model for the anomaly detection application: 1) our research can leverage on the large body of research on identification of ARIMA models developed over the past 50 years; 2) decentralized and multivariate extensions of recursive FARIMA model identification algorithms appears to be more straightforward; 3) as FARIMA models are defined explicitly as temporal state recursions, they would appear naturally suited to detecting transient disruptions; 4) causal and time recursive state and parameter estimation algorithms might more easily be developed.

Research issues? Inputs from Rich and Rolf?

Particle Filtering Models

Non-parametric Bayes classifiers based on predictive densities and sequential importance sampling provide another approach which is more computationally intensive but is optimal with respect to a performance criterion. When used with predictive densities such approaches fall somewhere between the model based and learning-based methods. A principal issue to be investigated is the development of distributed methods for Gibbs sampling and predictive density generation.

Need more from Mark???

4.6 Decentralized Pattern Recognition and Anomaly Detection

While pattern recognition will used to identify baseline and states, anomaly detection will be used to detect deviations from the baseline. The end goal is to perform event detection in real time. The inter-site communications burden may make it difficult to perform these operations at a central processing location. We will investigate decentralized methods, where clusters of data collection sites process locally while exchanging limited amounts of information. However, previously developed decentralized sequential change point detection algorithms [50, 42, 15] must be extended to the case of event-driven detection. It has been shown by our collaborators that the DEDS framework is well suited to decentralized failure diagnosis [**?**]. The challenge will be to extend this to the case of distributed anomaly detection. We will also investigate fully Bayesian decentralized detection methods using the sequential Monte Carlo Markov Chain framework such as that used in [11].

Demos?

4.7 Distributed Compression

Sandeep?

5. Data Collection (Leader: Prakash)

Data will be collected from a combination of cooperative sites and non-cooperative sites in the network. Cooperative sites will include routers, switches and terminals in Internet which are part of our consortium of sites. Data from the rest of the network, the non-cooperative sites, 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.

5.1: Consortium Sites:

Router Sites: Router data including netflow and timing will be collected in collaboration with Internet2 and Merit . All collected data will be rendered anonymous and raw data will destroyed within 72 hours in accordance with privacy protocols already in place at these collaborating facilities. Merit will provide the following services to the project:

1) Provide access to Netflow or similar data gathered from MichNet, the regional Internet network that Merit operates in Michigan;

2) Provide the project with space for at least one system in an equipment rack at Merit's Arbor Lakes data center;

3) Provide 24 hour a day 7 days a week monitoring of the system installed in the Arbor Lakes Data Center and its access to MichNet by Merit's Network Operations Center (NOC);

4) Provide assistance in the on-going operation of the system that will be installed in the Arbor Lakes Data Center;

5) Assist other project staff in network event characterization; and

6) 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.

7) Assign technical and management staff to preform the above tasks and to participate as active members of the project, advise on data collection strategies, help test DDoS and intrusion detection strategies, supervise student interns, and participate on the project's national advisory board.

Volunteer Sites: We will also develop an open collaborative infrastructure for data collection and distributed pattern recognition using a network of participating ip terminals and servers. This will involve free and wide dissemination of open-source software tools for data collection, computation, and transmission to Rice and UM. Similarly to the `SETI@home` program the software would take advantage of a participant's free cpu cycles to perform data collection and computations. The software would collect packet-header, timing, and other data from ip terminals. This data will supplement netflow and other data obtained from Merit and Internet2 routers to detect attacks involving multiple-site intrusions. A possible scheme is to allow the user to specify the level of collaboration but allocating varying amount of resources (disk space) through specification of a finite memory window over which data is stored, processed and forwarded. While any particular site may not collect data continuously over any long time interval the aggregate of all collaborating sites will produce a patchwork of asynchronous snapshots of the network at different sites and different times that will be sewn together and analyzed at UM and Rice. All collaborating sites will time stamp their data according to NIST/GMT time-clock information which will be downloaded periodically. Using such a volunteer data-collection network avoids many data-collection privacy issues as the participants are freely offering their data to the project. It also has the potential to grow organically to an unprecedented large-scale network of participating data-collection sites.

5.2: End-to-end Probing:

Measurements from only a few distributed nodes will be used to do change point detection, localization, and model identificaton as discussed above. These measurements will be composed of a mix of ambient traffic measurements and active probe measurements. This gives our system the flexibility of nominally monitoring only a few ambient flows for anomalies while bringing in more intrusive methods of active tomography only when these nominal measurements leads one to suspect that some significant deviation from baseline has occurred. Several members of our team have shown that unicast active probing can provide accurate estimates of link delay and loss characteristics from a few edge measurements without any special cooperation of routers in the network [9, 10, 48, 12]. Recently these methods have been extended to cases where traffic patterns may be changing over the probing period [11]. The challenge will be to integrate tomography methods into the kernel-based and DEDS modeling frameworks that we proposed above. A principal question that we will study will be how to effectively merge passive router measurements and active probing measurements at the data-aggregation centers for anomaly detection?

6. Applications (Leader: Liu)

The pattern recognition and detection framework described in the previous sections can potentially be applied to the early detection, prevention, and analysis of a variety of known attacks. It's underlying mechanism can also be used in accomplishing tasks like service monitoring that are critical to a wide range of applications. In this section we describe in more detail example applications that will be studied in this project, their relationship with the more general detection framework, and our proposed approaches. In particular, we we will discuss the applicability of our framework to the general class of distributed denial of service attacks, an example of which is the SYN flood attack. We will also discuss applications including service monitoring/verification for server applications, and network maintenance and provisioning.

DDoS attacks: Distributed denial of service (DDoS) attacks are an increasingly damaging class of attacks which can bring down servers via sending malicious packets at very high rate/intensity over a period of time. Thus DDoS is typically accompanied by significant and simultaneous changes in traffic-level and packet-level statistics at different locations. 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. To a large extent, all these approaches 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 on 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 on the attack is detected, the earlier actions can be taken and the network is better 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 in detecting such attacks. Another difficulty is that accurate detection of subtle changes is inherently subject to high false alarm rate. The proposed detection framework focuses on quickly distinguishing the (fast) spatio and temporal changes in traffic pattern via our novel combination of statistical multivariate models and DEDS models. We will evaluate the responsiveness, as well as accuracy of our system under attacks of different scale. How to enhance the responsiveness but at the same time decrease the probability of false alarms will be addressed by the proposed research. We will also conduct extensive experiments in measuring the level of service degradation between the initiation of these attacks and the detection and action launched by the proposed framework. Should these models prove useful, they can be retrofitted to any global infrastructure for combatting DDoS attacks.

One effective type of DoS attack can be achieved through a "SYN flood" [13], which consists of a stream of TCP SYN packets directed to a listening TCP port at the victim. Such a mechanism can be rendered extremely powerful, if it can be used from a set of compromised Internet nodes, where attack deamons 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, which is overwhelmed by the various requests. However, a key feature of this mechanism is that the attack programs select source IP addresses at random, or in statistical terms from a uniform distribution on the set of all IP addresses on the network.

Using a very simple *filtering* mechanism and the above uniform source address hypothesis Moore et al. [17] are able to identify DoS activity on the Internet. Their filtering mechanism is based on identifying packet flows with particular characteristics consistent with DoS attack mechanisms. Two significant shortcomings of their *backscatter* analysis are: (i) the uniform source address hypothesis, (ii) the fixed classification rules used and (iii) the ex-post nature of their methodology. We propose to extend this methodology in several new directions:

1. Examine the nature of multivariate backscatter distributions, both in time and in space; for example, under a random spoofing mechanism we expect to see highly correlated distributions over separate time windows. Examining backscatter distributions over space would also allow us to identify "reflector attacks" [17].

2. By using sequential statistical techniques [49] (e.g. sequential probability ratio tests, cumulative sum tests, etc) which have a well developed theory and possess a low computational complexity,

we would be able to detect *online* DoS attacks employing the above described mechanism.

3. Finally, by characterizing the backscatter multivariate temporal and spatial distributions, we could develop a database that would include samples of such distributions under various ”normal” network conditions, as well as under ”attack” conditions. Based on this *training* database, we would develop online classifiers that would be able to identify random source type DoS attacks.

It is worth noting that the above outlined methodology, namely characterize distributions of events, build a database of such distributions under different network conditions and finally develop online classifiers, could be used in several other settings, thus enabling us to detect and localize various types of anomalous network behavior.

Service monitoring/verification: 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 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, when a customer communicates via the Internet often a significant portion of the network connection is not under the direct control and responsibility of the service provider. If the customer experiences poor performance, it is difficult to determine whether it is due to the service provider's portion of the connection or the Internet at large. The only way to separate these effects and verify service is to assemble network performance information that is localized to the service provider's network. Directly measuring local network performance is very expensive for service providers. Furthermore, it is impossible for customers or third-parties to independently collect such information. End-to-end probing combined with network tomography makes it possible for an independent party to verify the service level without the cooperation of the provider and provides a cost-effective means by which the service provider can track the performance of their system.

Maintenance and Provisioning: Maintenance of a network is a major overhead, and when the performance of a network is poor, it can be very difficult to isolate the cause of the problem. Sometimes a router is performing sub-optimally; on other occasions, too much network traffic is directed along one path while other paths remain idle. Furthermore, as networks grow in complexity it is often the case that network owners may not be aware of all components in their system, and consequently methods for mapping the topology of a network are critical. The pattern recognition and detection methods developed in this project could be used to determine the topology and connectivity of the network, to localize poor performance to individual network components, to rapidly identify faulty components and to optimize routing decisions. The overall benefit is that maintenance overhead is significantly reduced. In addition, the proposed methodologies can indicate where additional network resources are required, which is critical for service provisioning.

7. Education (Leader: Nobel)

- Global network security curriculum (Nobel)
- Networking security internships (Liu)
- Networking laboratory (Liu, Nobel, Wallach, Prakash)
- Webcast seminars and workshops (Wallach)
- Summer UG and high school program (Hero)
- Network gaming competition (Prakash)

(*Liu, Prakash, Wallach, Noble, Baraniuk, Hero*)

The scope of this effort will provide many opportunities for undergraduate and graduate students to be involved in research. We also plan to include precollege students in this project through a summer internship program with private and public schools in Houston and Metro Detroit.

We are committed to making an impact on education for which we propose the following.

1. College students will be involved in developing a software tool for visualization of parameters in the spatio-temporal model and connectivity from the network tomography software.
2. A summer internship program will be developed to expose high school students to the areas of computer security and signal processing. This program will be coordinated with the existing high school program called Camp CAEN (Computer Aided Engineering Network) at UM.
3. One of the co-PIs has developed a random topology generator called Inet [33] that has been used by researchers in the networking field to generate realistic inter-domain topologies of the Internet. College students will be involved in testing out our anomaly detection algorithms on simulated networks.
4. We will develop a networking laboratory at Rice and UM which will serve the dual purpose of educating students in network security and providing data traces for our research. For more details on this lab see the budget justification. 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. A competition will be held between Rice University and University of Michigan during the summer of each year with prizes going to those students who develop the best attack strategy and the best thwarting strategy. Judges will be drawn from the team of co-PI's and collaborators. High school students will also participate in this competition through the summer internship program.
5. Undergraduate students will help acquire and analyze real network data at Rice and UM in the context of classroom instructional laboratories and independent study projects.

The proposed research will also help us improve our curriculum in networking and security. We have recently made senior design a requirement for our degree programs. The proposed research should help us define senior design projects in networking and security courses at the undergraduate level. At the graduate level, co-PI Prakash is planning to teach a pilot course on network security in Winter 2002, which we plan to make a regular course. Intel has recently donated 25 laptops to co-PI Prakash to allow ad hoc networks to be set up for projects related to networking and security graduate curriculum.

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] S. Axelsson. *Intrusion Detection Systems: A Survey and Taxonomy*.

[3] A. Beneviste, E. Fabre, and et al. *Diagnosis of Asynchronous Discrete Event Systems, A Net Unfolding Approach*. Preprint.

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

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

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

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

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

[9] M. Coates and R. Nowak. *Network Loss Inference using Unicast End-to-end measurement*, Sep. 2000.

[10] 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.

[11] 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.

[12] 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`.

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

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

[15] R. W. Crow, "Quickest detection for sequential decentralized decision systems," *IEEE on Aerospace Electronics Systems*, vol. AES-32, no. 1, pp. 267–283, Jan. 1996.

[16] 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.

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

[18] 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.

[19] J. A. Fessler and A. O. Hero, "New complete-data spaces and faster algorithms for penalized-likelihood emission tomography," in *Proc. of IEEE Nuclear Science Symposium and Medical Imaging Conf*, pp. 1897–1901, San Francisco, November 1993.

[20] 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.

[21] 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.

[22] J. A. Fessler and A. O. Hero, “Complete data spaces and generalized EM algorithms,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, pp. IV.1–IV–4, Minneapolis, MN, April 1993.

[23] J. Fessler and A. Hero, “Cramer-Rao bounds for biased estimators in image restoration,” in *Proc. of 36th IEEE Midwest Symposium on Circuits and Systems*, Detroit, MI, Aug. 1993.

[24] A. O. Hero, “The influence of the choice of complete data on convergence of EM-type algorithms,” in *Proc. of the IEEE Workshop on Statistical Signal and Array Processing*, pp. 74–77, Victoria, Oct. 1992.

[25] A. O. Hero, “Theoretical limits for optical position estimation using imaging arrays,” in *Actes du Colloque GRETSI*, pp. 793–796, Juan-les-Pins, France, Sept. 1991.

[26] 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.

[27] 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.

[28] A. O. Hero and J. A. Fessler, “A fast recursive algorithm for computing CR-type bounds for image reconstruction problems,” in *Proc. of IEEE Nuclear Science Symposium*, pp. 1188–1190, Orlando, FA, Oct. 1992.

[29] 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`.

[30] 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.

[31] 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.

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

[33] C. Jin, Q. Chen, and S. Jamin, “Inet: Internet topology generator,” Technical Report CSE-TR-433-00, University of Michigan, EECS Dept., 2000. http://topology.eecs.umich.edu/inet.

[34] 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.

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

[36] S. Lafortune, D. Teneketzis, M. Sampath, R. Sengupta, and K. Sinnamohideen, “Failure diagnosis of dynamic systems: An approach based on discrete event systems,”, pp. 2058–2071, June 2001.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] A. R. Reibman and L. W. Nolte, “Optimal design and performance of distributed signal detection systems with faults,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 10, pp. 1771, 1990.

[43] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk, "Multifractal cross-traffic estimation," in *Proceedings of the ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey, CA, Sept. 18-20 2000.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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`.

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

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

[51] M. Usman and A. Hero, "Recursive CR-bounds: algebraic and statistical acceleration," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Adelaide, Australia, April 1994.

[52] M. Usman, A. Hero, and J. A. Fessler, "Uniform CR bound: implementation issues and applications to image reconstruction," in *Proc. of IEEE Nuclear Science Symposium and Medical Imaging Conf*, pp. 533–537, Virginia Beach, November 1994.

[53] M. Usman, A. Hero, J. A. Fessler, and W. Rogers, "Bias-variance tradeoffs analysis using uniform CR bound for a SPECT system," in *Proc. of IEEE Nuclear Science Symposium and Medical Imaging Conf*, pp. 1463–1467, San Francisco, November 1993.

[54] M. Usman, A. Hero, and W. L. Rogers, "Performance gain analysis for adding a vertex view to standard SPECT," in *Proc. of 36th IEEE Midwest Symposium on Circuits and Systems*, August 1993.

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

co-PI	Algorithms/Models	Data Collection	Applications	Education
A. Hero(UM)	X		X	
R. Baraniuk(Rice)	X			X
A. Benveniste(INRIA)	X		X	
M. Coates(McGill)	X		X	
R. Dwarshuis(Merit)		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.

8. Management Plan

To accomplish the research and education aims of this project requires a *focused large scale and multi-disciplinary effort*. This UM/Rice/Merit /Arbor team brings complementary strengths to this project necessary to make a leap forward in detection of network anomalies. We illustrate the commonalities and differences of our various strengths in the Venn diagram on Figure 1. The breakdown of senior personnel by sub-areas in Table 2 illustrates the balance of our team.

The coordination of collaborative projects which span three colleges (UM, Rice, McGill, INRIA-Rennes) and two networking organizations (Internet2 and Merit) requires a tight management plan. The challenge is heightened by the fact that the team crosses traditional boundaries which have separated networking and computer security from signal processing, control and statistics. Research partnerships which lead to productive inter-disciplinary collaborations will be essential for success of this effort. To facilitate and enhance such collaborations each co-PI will been allocated one and a half graduate students research assistants (GSRA) where the half GSRA will be co-supervised by two co-PI's each coming from a different area (networking, statistics, signal processing, and control systems). The allocation of 1.5 GSRA's per co-PI will allow each co-PI to pursue a collaborative research project in addition to a focussed core research topic within the co-PI's specialty area. In addition we will develop and team-teach courses combining signal processing, control systems, statistics, and network security. Furthermore, close collaboration with collaborators and co-PI's at Internet2, Merit , and Arbor Networks on solving real-life networking security problems will help provide the focus necessary to integrate the diverse expertise represented by our team.

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 impact of the project. We have agreements with Sprint, Stanford SLAC, Internet2, Lucent Bell Labs (Murray Hill), Merit Networks, Inc., and Arbor Networks to serve on the NAC (see attached letters). 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 (NAC) will be created with representatives from the university computing organizations and legal counsel. The equivalent of a .25% legal assistant (JD) has been retained to advise us on what will undoubtedly be considerable privacy issues associated with collecting and disseminating sanitized data from the HeaderVault and the switch monitoring sites deployed at UM and Rice.

Electronic Dissemination: A website will be created to archive 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 much 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.

High School Summer Camp in Network Security: A summer program for high school students will be organized every year. This program will consist of an intensive two week camp where they will learn about computer networks in a hands on educational and recreational environment. We will seek out kids from a variety of backgrounds, including under-represented socio-economic groups in Houston and Detroit, to participate. We will work closely with Camp CAEN (`http://campcaen.engin.umich.edu/`), a computer exploration summer camp at the College of Engineering at UM, to recruit high school students and to develop a fun-filled network security camp curriculum.

Residency program and minisymposia: Each year we will run a 2 week residency program in network security. 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 is 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. University of Michigan facilites (Cambridge House) 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. As contrasted to the short courses, which are aimed at continuing education of networking professionals, the workshops will be aimed toward the academic community. The workshops will take place over three successive days. They will 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.

Human Subjects and Internal Review Board (IRB)

The raw data collected by Merit and CITI includes header information and other potentially sensitive private information that identify IP source and other information. Our data collection collaborators at Merit networks and CITI will anonomize the data. Safeguarding the privacy of the data traces will be of primary concern in this project and we have included a legal assistant (JD) to copunsel us on these matters. Issues of privacy and human subjects will be addressed with the IRB if this proposal is funded.

Team Management:

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 respoective 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.
- Identifying potential difficulties impeding progress

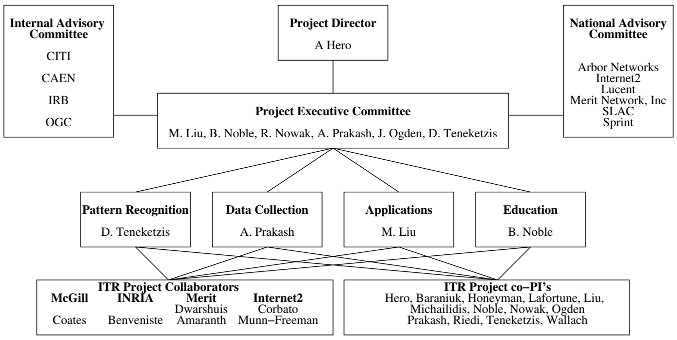

Figure 3: *Management structure for the project. Executive committee will operate as the clearinghouse for all decisions and will solicit inputs from the National Advisory Committee (NAC) - whose members are indicated - and an Internal Advisory Committee (IAC) which will be composed of representatives 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.*

- 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 quarterly 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 National Advisory Committee (NAC) once per year.

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, etc).

A time line for the five year duration of the project is given in Tables **??** and **??**. The first year and the last year we will have face-to-face kick-off and wrap-up meetings involving all co-PI's, collaborators, and the NAC. Other industry and government representatives will be invited to attend this meeting. In addition to these larger meetings there will be several meetings over each year of the project. These will include monthly meetings via webcast/videoconference/teleconference to assess progress and explore new ideas for education and research. Several UM co-PI's will visit co-PI's at Rice, and vice versa, twice a year. One or two day year-end meetings of all co-PI's and collaborators will take place in Ann Arbor. At this meeting co-PI's and collaborators will present previous year's research and education results to the National Advisory Committee. Other industry and government representatives will be also invited to participate. At these meetings we will set or refine specific goals for the next year.

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
♦	Global Network Security UG/G course
◎	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 2)

Budget Justification

1. We are including Prof. Mark Coates and Prof Albert Benveniste on this grant as international collaborations (see attached letters). Dr. Coates is an Assistant Professor at McGill University in Montréal Canada in Winter 2002. He has been a research collaborator with Profs. Nowak and Hero and his expertise in stochastic optimization using Monte Carlo markov chains and sequential importance sampling will be crucial to the pattern recognition efforts. Prof. Benveniste is a Professor at the University of Rennes and the Director of Research at INRIA (IRISA) in Rennes, France. Prof Benveniste is a specialist in network failure detection and isolation and will play a key role in adapting the DEDS framework to the anomaly detection problem. No ITR funds are requested for Dr. Coates or Dr. Benveniste. However, if this ITR is funded we will apply for additional support for their efforts in the NSF Office of International Programs.

2. Funding for two post-docs are requested for this project. This will provide post-graduate education to the post-docs and full time research assistance to co-PI's.

3. $?K is requested for equipment to set up isolated network environments at Rice and UM to be used for educational of college and precollege students on network security, generate attack scenarios in the context of the summer gaming competition, and serve as a small scale testbed for co-PI's and their graduate students. The networks will consist of two LAN's each consisting of 50-75 PC terminals and routers (both wireless and wired) at Rice and UM. Communications between the LAN's at Rice and UM will be through the Internet using a pair of dedicated DNS servers.

4. $500K is requested for two full time lab adminstrators to help develop and maintain the lab environments at Rice and UM over the five years of the grant.

5. $30K per year is requested to fund the summer camp program for high school students. This will be enough to set up the program each year (recruiting, publicity, etc) and provide scholarships to 20-30 students to attend Camp CAEN from Rice, UM and elsewhere.

6. We request $30K per year for funding undergraduate students to help develop the network laboratory environment, to help develop web tools, and to help analyze data traces. Supplementary funding for undergraduate research projects will also be requested from NSF.

7. $30K/year for a 50us develop the downloadable software for the volunteer data collection network.

8. $20K per year is requested for UM PI and co-PI travel (Rice co-PI travel is bundled into Rice subcontract).

9. $50K per year is requested for support of the summer residency program. This will cover travel expenses of the 2 organizers and 20 invited participants in addition to a per diem for expenses incurred during the two week long stays of the participants.

10. Equipment: 10 Header-vault sniffer systems (a PC with a fast CD-ROM read/writer for archiving) will be deployed at a cost of $2000 each. 2 identical router sniffer systems each costing $5000 to be installed at Merit and at Rice (excludes space rental rental fee and administration fee). 15 workstations (one per student) dedicated to the project for code development and testing, visualization, and trace analysis ($2000 each to be installed in EECS building). 2 high-end workstations ($5000 each for dual-processor 2GHz machines with 2GB memory) to perform computation-intensive monte carlo and other simulations.

11. $25K per year is requested for computer security course development and web-delivery to K12 teachers and computer technical support staff at K12 schools throughout Michigan. This activity will be administered by Merit staff-member J. Ogden in collaboration with UM faculty co-PI's.