Quantitative Methods in Defense & National Security

QMDNS 2007

George Mason University
February 7-8, 2007

Statistics in Defense & National Security
SDNS

Sponsored by American Statistical Association, Section on Statistics in Defense and National Security, Interface Foundation of North America, and Washington Statistical Society

Quantitative Methods in Defense and National Security 2007

George Mason University
February 7-8, 2007

Program Chair

Dr. Edward J. Wegman
College of Science
George Mason University
4400 University Drive
Fairfax, VA 22030-4422
ewegman@galaxy.gmu.edu

Technical Program Chair

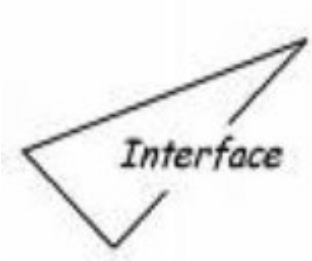
Dr. Jeffrey Solka
Dahlgren Division of the Naval Surface Warfare Center
Principal Scientist
Electromagnetic Systems and Sensors Q Department
Staff Scientist Advanced Science and Technology Division Code Q20
17320 Dahlgren, Rd
Dahlgren VA, 22448-5100
Jeffrey.Solka@navy.mil

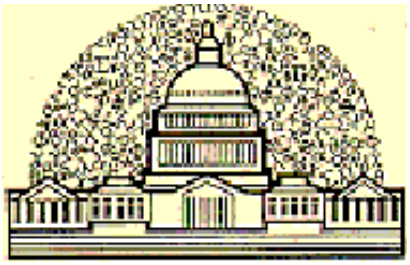
Keynote Address

Dr. Michael McGrath
Deputy Assistant Secretary of the Navy (Research, Development, Test & Evaluation)

Sponsors

American Statistical Association
Section on Statistics in Defense and National Security
Interface Foundation of North America
Washington Statistical Society





Wednesday, 7 February 2007

	Track A - Johnson Center, Dewberry Hall North	**Track B** - Johnson Center, Dewberry Hall South
8:00-8:30	Conference Registration: Johnson Center, Dewberry Hall Lobby	
8:30-10:00	**Plenary Session:** Johnson Center Cinema Keynote Address: **Dr. Michael McGrath, Department of the Navy**	
10:00-10:30	Break	
10:30-12:00	**Invited Session #1:** "*Homeland Security Applications"* Organizer and Chair: Suzanne Strohl (DHS) • Guy Thomas, Coast Guard • Mark Andress, Office of Naval Intelligence	**Invited Session #2:** *"Spatial and Time Series Analysis"* Organizer and Chair: Myron Katzoff (CDC) • S.N. Lahiri, Texas A & M University • Carol A. Gotway Crawford, Centers for Disease Control and Prevention • Nozer D. Singpurwalla, The George Washington University
12:00-1:30	Lunch	
1:30-3:00	**Invited Session #3:** *"Intelligence Applications"* Organizer and Chair: Helen Gigley (U.S. Govt.) • Steven Rieber, Georgia State University • Daniel B. Carr, George Mason University	**Contributed Session #1:** *"Modeling"* Chair: Elizabeth Hohman • Michael B. Hurley, MIT Lincoln Laboratory • Dr. Bill Luker Jr., Lockheed Martin IS&S • Michael J. Kwinn, Jr., United States Military Academy
3:00-3:30	Break	
3:30-5:00	**Contributed Session #2:** *"Biosurveillance - I"* Chair: Brandon Higgs • Robert J. Gallop, West Chester University • Jaideep Ray, Sandia National Laboratories • Arvind K. Jain, RAND Corporation	**Contributed Session #3:** *"Data Analysis – I"* Chair: Nicholas Tucey • Kai-Sheng Song, Florida State University • John M. Conroy, IDA Center for Computing Science • Aubrey Magoun, University of Louisiana at Monroe
6:00-8:00	**Poster Session and Reception:** Concert Hall 3rd Floor Lobby	

Thursday, 8 February 2007

	Track C - Johnson Center, Dewberry Hall North	**Track D** - Johnson Center, Dewberry Hall South
8:00-8:30	Conference Registration: Johnson Center, Dewberry Hall Lobby	
8:30-10:00	**Topic Contributed Session#1:** *"Data Fusion"* Organizer: Ed Wright Chair: Wendy Martinez • Ed Wright, IET, Inc. • Jim Jones, SAIC and Ferris State University • Kathryn Blackmond Laskey, George Mason University	**Topic Contributed Session #2:** *"Text Data Mining"* Organizer and Chair: Jeffrey Solka • Jeffrey Solka, NSWC • Elizabeth Hohman, NSWC • John Rigsby, NSWC
10:00-10:30	Break	
10:30-12:00	**Topic Contributed Session #3:** *"Network Analysis"* Organizer: David Marchette Chair: John Rigsby • David Marchette, NSWC • John E. Gray, NSWC • Leslie McIntosh, Saint Louis University	**Topic Contributed Session #4:** *"Operations Research"* Organizer and Chair: Don Wagner • Moises Sudit, University at Buffalo (SUNY) • Shanchieh Jay Yang, Rochester Institute of Technology • L. D. Servi, MIT Lincoln Laboratory
12:00-1:30	Lunch	
1:30-3:00	**Contributed Session #4:** *"Biosurveillance - II"* Chair: Chris Overall • William R. Hogan, University of Pittsburgh • Howard S. Burkom, The Johns Hopkins University Applied Physics Laboratory • Gerald Shoultz, Grand Valley State University	**Contributed Session #5:** *"Data Analysis - II"* Chair: David Marchette • Jonathon Phillips , NIST • Pranab K. Banerjee, Space Dynamics Laboratory • Patricia H. Carter, NSWC
3:00-3:30	Break	
3:30-5:00	**Contributed Session #6:** *"Survey and Modeling"* Chair: Yasmin Said • David Banks, Duke University • Paul B. Massell, U.S. Census Bureau • Genetha Gray, Sandia National Labs	**Contributed Session #7:** *"Anomaly Detection"* Chair: Barton Clark • Michael D. Porter, NCSU • Deepak Agrawal, Yahoo! Research • Christopher Overall, George Mason University

Directions to George Mason University, Fairfax Campus

from I-95 (north and south)

From points north on I-95, take exit 27 (I-495 West), then follow the directions below "from the Capital Beltway (I-495)." From points south on I-95, take exit 160B (Route 123 North). Follow Route 123 for approximately 15 miles to Braddock Road. Turn right on Braddock Road. At the first signal, turn left on Roanoke River Road. Bear right at the fork in the road. Take the first left on Mason Pond Drive to the Mason Pond Parking Deck, the last building on your right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

from the Capital Beltway (I-495)

Take exit 54, Braddock Road (Route 620) West. Follow Braddock Road for approximately six miles. Pass the first entrance to the university and turn right at the stop light at Roanoke River Road. Bear right at the fork in the road. Take your first left onto Mason Pond Drive; parking is available in the Mason Pond Parking Deck, the last building on the right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

via I-66E from Front Royal (Fairfax County Parkway)

Exit at the Fairfax County Parkway South (Route 7100). Exit the parkway at Braddock Road, stay to the left, and turn left onto Braddock Road. Take the first left past Route 123 (Ox Road) onto Roanoke River Road. Bear right at the fork in the road. Take the first left on Mason Pond Drive to the Mason Pond Parking Deck, the last building on your right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

via I-66W from Washington, D.C., or Arlington

Take exit 60 at Route 123 South, Chain Bridge Road. Follow Route 123 through the City of Fairfax, and turn left at University Drive. Take the first right at Occoquan River Lane. Turn right at the stop sign onto Patriot Circle. At the pond, bear left to stay on Patriot Circle. Take the first left on Mason Pond Drive to the Mason Pond Parking Deck, the last building on your right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

via Route 50 from Washington, D.C., or Arlington

Take Route 50 West to I-495 South (Capital Beltway) toward Richmond. Take exit 54, Braddock Road (Route 620) West. Follow Braddock Road for approximately six miles. Pass the first entrance to the university and turn right at the next light at Roanoke River Road. Bear right at the fork in the road. Turn left on Mason Pond Drive; parking is available in the Mason Pond Parking Deck, the last building on the right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

from Ronald Reagan National Airport

When exiting the airport, follow the signs to Washington, D.C., North and to the George Washington Parkway North, to I-395. Once on the parkway, stay in the middle lane to enter I-395 South to Richmond. Immediately move left three lanes to remain on I-395 South. Exit I-395 at I-495 North (exit 1C) to Rockville. Exit I-495 at exit 54, Braddock Road West.
Follow Braddock Road for approximately six miles. Pass the first entrance to the university and turn right at the next light at Roanoke River Road. Bear right at the fork in the road. Turn left on Mason Pond Drive; parking is available in the Mason Pond Parking Deck, the last building on the right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

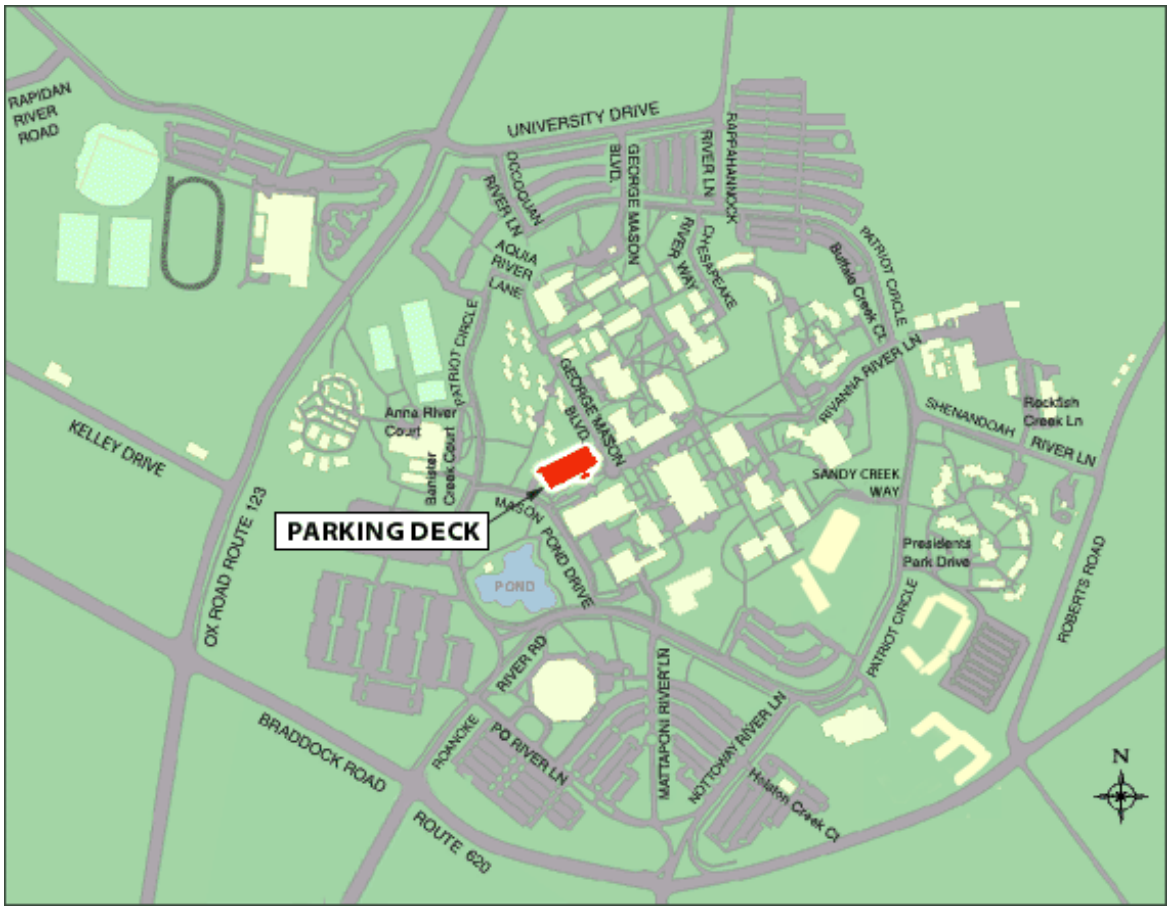
from the Vienna Metro Station

Take the CUE Bus (any of the routes stop at George Mason). The CUE Bus is free with a valid Mason identification card, or is $.50 otherwise. If traveling by car, exit the Vienna Metro on Nutley Avenue South (Route 243). Follow Nutley Avenue to Route 50 (Arlington Boulevard) and turn right. Stay on Route 50 to Route 123 (Chain Bridge Road) and turn left. Follow Route 123 through the City of Fairfax to University Drive, then turn left. Take the first right at Occoquan River Lane. Turn right at the stop sign onto Patriot Circle. At the pond, bear left to stay on Patriot Circle. Take the first left on Mason Pond Drive to the Mason Pond Parking Deck, the last building on your right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

from Dulles Airport

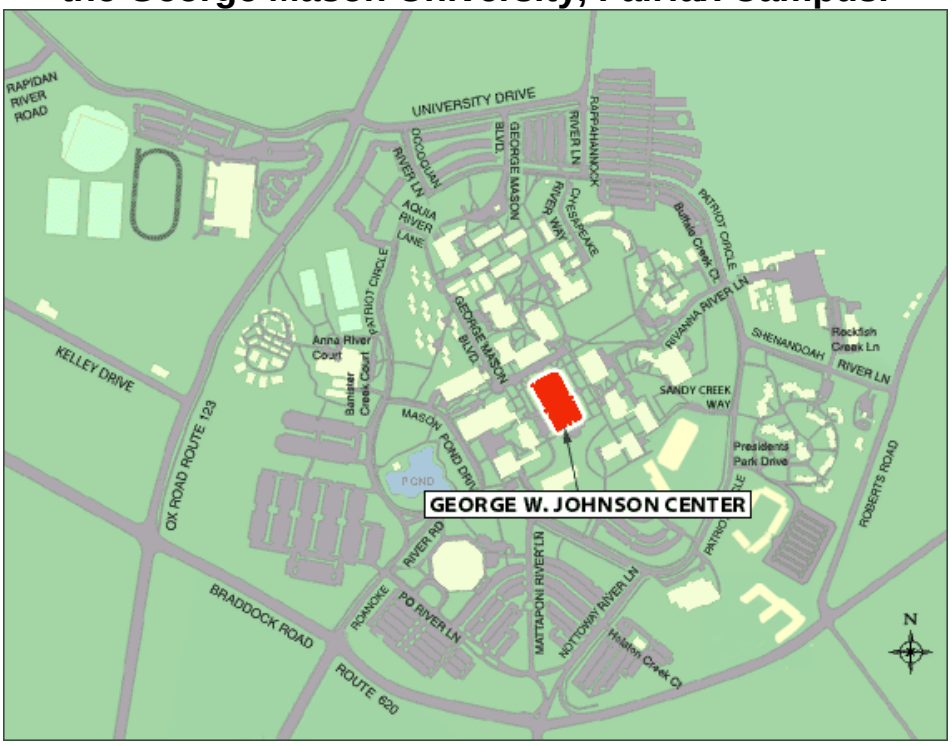
Exit the airport onto the Dulles Access Road, which leads to the Dulles Toll Road at Route 267, Reston Parkway. Exit the Dulles Toll Road (no toll required) onto Fairfax County Parkway (Route 7100). Exit the parkway at Braddock Road. Stay to the left, and turn left on Braddock Road. Take the first left past Route 123 onto Roanoke River Road, and go right at the fork in the road. Take your first left onto Mason Pond Drive and continue to the Mason Pond Parking Deck, the last building on your right. An information kiosk is located outside the third level of the deck to help you navigate the campus.

Parking at the George Mason University, Fairfax Campus

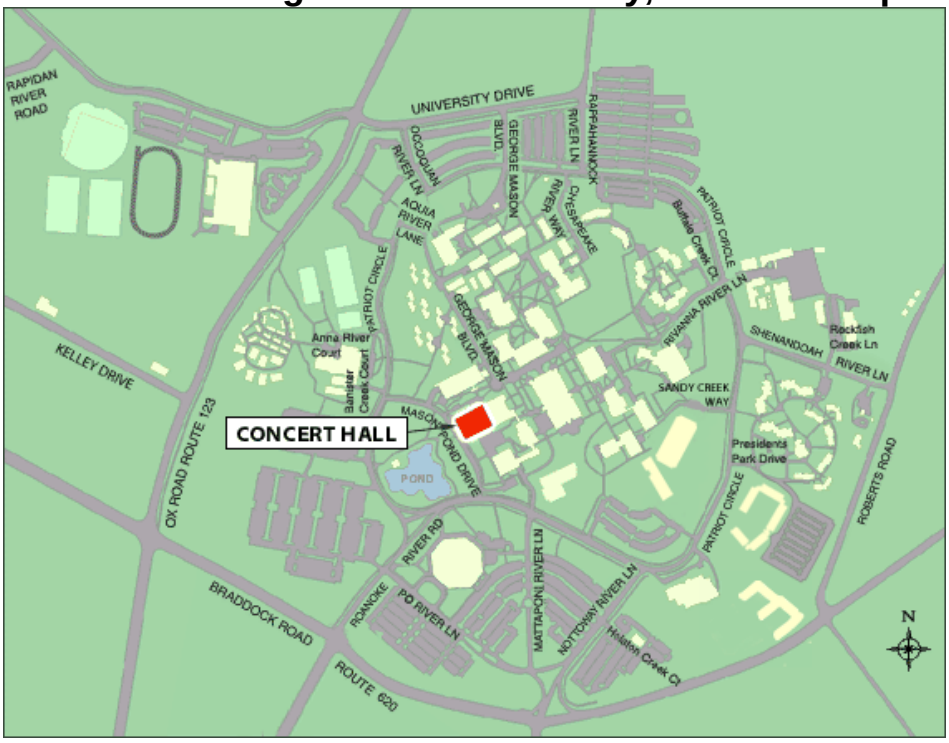


Our parking vouchers are only valid for this garage. Our vouchers do not work for the Sandy Creek Parking Garage.

The conference will be held in the George W. Johnson Center on the George Mason University, Fairfax Campus.



The Poster Session will be held on the third floor of the Concert Hall on the George Mason University, Fairfax Campus.



The conference sessions and registration will be held on the ground floor of the Johnson Center in Sid and Reva Dewberry Hall (G26). The Keynote Address will be in the Cinema (G30).

GROUND FLOOR

Advanced Internet Laboratory	G10
Bistro	G38
Catering Office	G18
Cinema	G30
Dance Studio	G34 & G35
Events Production	G45
Gold Room	G19
Loading Dock	
Meeting Room H	G19
Jazzman's	G32
Sid & Reva Dewberry Hall	G26
WGMU	G33

George W. Johnson Center Directory

Accessible to Disabled
Automatic Teller Machine
Elevator
Food Service
Information
Northernly Direction
Men's Restroom
Public Lockers
Public Phones
Security
Stairway
Women's Restroom

Full Schedule and Abstracts

Wednesday, 7 February 2007

Wednesday, 8:30 – 10:00 Keynote Address

Keynote Address
Dr. Michael McGrath,
Deputy Assistant Secretary of the Navy (Research, Development, Test & Evaluation)

Abstract: Statistics are a central component in a variety of analytical disciplines that support Navy and Marine Corps programs and operations. Dr. McGrath will discuss the impact of statistical analysis on biometrics research and development programs and how statistical analysis within supercomputing supports operational planning, and ship and aircraft design and testing within the Navy and Marine Corps.

Bio:: Dr. Michael McGrath is the Deputy Assistant Secretary of the Navy for Research, Development, Test and Evaluation. His role is to aggressively drive new technologies from all sources across Navy and Marine Corps platforms and systems, and to develop programs to bridge the gap in transitioning from Science and Technology to Acquisition.

Prior to his appointment to this position in 2003, Dr. McGrath spent five years as Vice President for Government Business at the Sarnoff Corporation, a leading R&D company with both commercial and government clients.

Dr. McGrath has 28 years of prior government experience, in reverse life cycle order. He started in weapon system logistics at NAVAIR in the 1970s, moved into acquisition in the Office of the Secretary of Defense in the 1980s, and then into technology development at DARPA in the 1990s.

Dr. McGrath holds a BS in Space Science and Applied Physics, an MS in Aerospace Engineering, and a doctorate in Operations Research from George Washington University.

Wednesday, 10:00 – 10:30	BREAK
Wednesday, 10:30 – 12:00	Track A - **Invited Session #1:** *"Homeland Security Applications"* Organizer and Chair: Suzanne Strohl (DHS)

Systems Engineering the MDA System -Using Capabilities Based Assessment to Develop and Implement the National Maritime Domain Awareness System Implementation Plan
Guy Thomas, (Coast Guard), George.G.Thomas@uscg.mil

Abstract: The National Strategy for Maritime Security, and its supporting National Plan to Achieve Maritime Domain Awareness, written to fulfill the requirement of the joint National and Homeland Security Presidential Directive, NSPD-41/HSPD-13, Maritime Security Policy reflects the maritime security challenges of the 21st Century and directs the development of a sustained, continuous collaborative effort across the entire government of the United States, working with state and local governments, private organizations and foreign partners to provide a complete system to develop critical intelligence and information from all sources, classified and open source, and provide it to maritime operational commanders at all levels in a clear and concise manner in time to allow them to make the correct operational decisions and initiate the correct tactical action. It is clearly recognized by all participants in this development effort that successful MDA Implementation Plan, and its attendant Investment Strategy, execution demands unprecedented cooperation and information sharing among governmental agencies and organizations at all levels as well as the maritime industry, and international partners. The system requires an enhanced collaborative information environment (CIE) made up of information from human intelligence collection, defense, law enforcement and private organizations, and the integration of existing and emerging sensor technologies, analyzed and fused in an operator user-definable common operating picture (UDOP) operating in a multi-level security environment. Users with the highest clearance level would have access to all information, with those at lower levels of security clearances only having access to information appropriate for their level of clearance. Provisions for special access programs would also need to be accommodated. The technology exists to build such a multi-level system, however, it is getting all parties to agree to build it, and develop a real CIE within the total community of interest (COI) is the real challenge. Getting all to agreed to change their policies, procedures and, in some cases, the governing laws, is the real problem. Indeed, we do need to improve many aspects of our technology, but that is the easy part; getting buy-in from all of the participants to such an extent that they are willing to return to their parent organizations and move to get policies, procedures and laws changed is the hard part. This multi-dimensional task requires an equally sophisticated implementation effort to ensure the requirements of all stakeholders are fully considered and maximum buy-in is achieved across all departments, agencies, and all other entities within and without of government. . Thus, even the development of the implementation plan itself calls for the use of proven systems engineering methodologies and a structured analytical approach to the multi-departmental, complex problem of building an effective national maritime domain awareness system.

Maritime Domain Awareness (MDA) Data Sharing (DS) Community of Interest (COI)
Mark Andress, (Office of Naval Intelligence), mandress@nmic.navy.mil,
Beth Gorko, (Coast Guard), beth.gorko@uscg.mil, and
John Macaluso, (Coast Guard), john.j.macaluso@uscg.mil

Abstract: The Maritime Domain Awareness (MDA) Data Sharing (DS) Community of Interest (COI) was established in February 2006 to focus on maritime information sharing among federal agencies and their partners. The primary objective of the MDA DS COI is to develop a data standard that supports net-centric information sharing across the full spectrum of MDA stakeholders, culminating in the visibility, accessibility, and understandability of data on a User Defined Operational Picture (UDOP).

The MDA DS COI is led by the Executive Committee, co-chaired by the U.S. Navy and U.S. Coast Guard. In addition to the Executive Committee, there is a GS15/O-6 Steering committee and three Working Groups. The Data Management working group is charged with developing the common vocabulary and

schema for the data sets selected; the Pilot Demonstration working group is charged with delivering a real-world technological demonstration that implements the common schema in a net-centric environment to share data among community members; and the Joint Implementation & Services working group is charged with identifying future Spirals and establishing relationships with potential partners.

There are multiple Spirals associated with the MDA DS COI, with each Spiral consisting of a Development Phase and Assessment Phase. The first Spiral provides a limited number of sources of unclassified Automatic Identification System (AIS) information net-centrically available on the Defense Information Services Agencys (DISA) Net-Centric Enterprise Services (NCES). Spiral 1 AIS Publishers include: U.S. Coast Guard Research and Development Center, Office of Naval Intelligence, Department of Transportation (DoT) VOLPE Center, and U.S. Navy Organic (shore/afloat).

Follow-on Spirals will increase AIS information available via NCES to COCOMS and non-DoD Intelligence and Operations centers (e.g., the National Operations Center (NOC)). In October 2006, the MDA DS COI Spiral 1 Pilot Flag Demonstration took place at the National Maritime Intelligence Center (NMIC). This event was very well attended and received very positive feedback from DoD and DHS stakeholders. The Pilot demonstrated a new paradigm for information sharing that decoupled the source providers from their normal display and provided the information to any available consumers (authenticated and authorized) in a common vocabulary using the DISA NCES messaging bus. The Pilot also demonstrated a new method for an unanticipated authorized user to first discover AIS data, then present the data on a given UDOP display (I-Map, Google Earth, etc.). The level of visibility of the COI work has increased significantly since the initial Pilot, with plans for another Executive Demonstration and the continuation of way ahead discussion focused on the scope, funding, and transition path of future MDA DS COI Spirals.

Wednesday, 10:30 – 12:00 Track B - **Invited Session #2:** *"Spatial and Time Series Analysis"*
Organizer and Chair: Myron Katzoff (CDC)

Spatial and time series methods for national security applications
S.N. Lahiri, (Texas A & M University), snlahiri@stat.tamu.edu

Abstract: In this talk, we present an overview of some statistical methodology for Spatial data and time series data that have direct applications to problems in national security. Specifically, we consider the problem of designing an optimal network for monitoring the (land) border through a combination of manned and wireless micro-sensors. The second problem we consider is the reconstruction/prediction of a multiple time series based on partial information obtained by sampling different components at different timepoints. The predicted time series can be used for developing a profile of an entity of interest (people, organizations,etc.)

Combining Incompatible Spatial Data
Carol A. Gotway Crawford, (Office of Workforce and Career Development, Centers for Disease Control and Prevention), cdg7@cdc.gov, and
Linda J. Young, (Department of Statistics, University of Florida), ljyoung@ufl.edu

Abstract: Many programs and studies increasingly use existing data from many different sources (e.g., surveillance systems, health registries, governmental agencies) for analysis and inference. More often than not, the data have been collected on different geographical or spatial units, and each of these may be different from the ones of interest. There are many statistical issues associated with combining such disparate data. This presentation provides an introductory overview of several such issues, including the problems that can occur when making inferences from aggregated data, the concept of spatial support, and the importance of proper uncertainty assessment. From this perspective, the utility of many different statistical approaches to the problem of combining incompatible spatial data will be assessed. We review the current statistical methods for combining incompatible spatial data including raster and geoprocessing GIS operations, centroid smoothing techniques, regression methods, multi-level tree models, Bayesian

models, and geostatistical methods. Several examples from public health will illustrate the relevant statistical issues.

The Utility of Survivability

Nozer D. Singpurwalla, (The George Washington University), nozer@gwu.edu

Abstract: The motivation for this research was a problem that the author confronted during some consulting work done, several years ago, for the Marine Corps. The general issue is, how does the DoD or other such agencies specify reliability requirements?. Numbers like .95 and .99 seem to be routine. Is there a justification for this choice? If so, the agency's utility should come into play. The problem therefore is an elicitation of utilities. But this is a general topic and should be of interest to statisticians as well as engineers doing control theory, and information integration, and decision makers in general. Statisticians and engineers use squared error and absolute error as a matter of routine. The matter of utility was of interest to statisticians like Savage, Chernoff, Rubin (Herman, that is), and Mosteller, and economists like Friedman, Arrow and Stigler, to name a few. Of late the matter seems to have become dormant. This talk is about peeking a t the old world through a new window. Modern decision theorists and economists write a lot about utilities. But the use of statistical ideas and methods seems to not have been exploited. We argue how this can be done.

Wednesday, 12:00 – 1:30 LUNCH

Wednesday, 1:30 – 3:00 Track A - **Invited Session #3:** *"Intelligence Applications"*
Organizer and Chair: Helen Gigley (U.S. Government)

Communicating Uncertainty in Intelligence Analysis

Steven Rieber, (Office of the Director of National Intelligence), stevedr0@dni.gov

Abstract: The current approach to conveying uncertainty in intelligence analysis is to qualify judgments with verbal expressions such as probably, unlikely, may and could, among others. No common standards govern the use of these terms within the Intelligence Community (IC). Although this approach gives analysts and managers maximal latitude, it risks serious and frequent miscommunication. Numerous studies have demonstrated wide variations in how people understand these terms. These diverse understandings mean that every time one of the terms is used, intelligence consumers might interpret it differently from how it was intended.

The potential miscommunication is not limited to interactions between analysts and consumers. When intelligence products are coordinated, analysts, offices, and agencies agree to common language. However, if the language itself is ambiguous in ways that are not evident to the coordinating parties, an illusion of agreement may mask very different understandings. In addition, intelligence products are typically read by several policy-makers, who may interpret terms such as possible differently from one another. Consequently ambiguity in these terms can also lead to misunderstandings between the policy-makers themselves.

The problem is not new. Sherman Kent recalled writing a 1951 National Intelligence Estimate (NIE) which stated that a Soviet attack on Yugoslavia should be considered a serious possibility. Shortly after the NIE was disseminated, the chairman of the State Department's Policy Planning Staff asked Kent what he meant by serious possibility. Kent writes:

I told him that my personal estimate was on the dark side, namely that the odds were around 65 to 35 in favor of an attack. He was somewhat jolted by this; he and his colleagues had read serious possibility to mean odds very considerably lower. Understandably troubled by this want of communication, I began asking my own colleagues on the Board of National Estimates what odds they had had in mind when they agreed to that wording. It was another jolt to find that each Board member had had somewhat different odds in mind and the low man was thinking of about 20 to 80, the high of 80 to 20. The rest ranged in between.

This type of miscommunication may be far more common than generally believed. Rarely do analysts follow Kent's example in asking consumers and other analysts how they understand probabilistic terms. Consequently, no one knows how often or how severely such terms are misinterpreted in practice.

Kent's proposed remedy was to standardize the meanings of a small number of qualitative expressions almost certain, probable, chances about even, probably not, and almost certainly not by stipulating that each should express a particular range of probabilities. For example, probable was to mean 75%, give or take about 12%. Despite sporadic attempts at standardization, Kent's proposal was never widely adopted, and the dilemma of how to minimize or eliminate miscommunication in intelligence products remains unsolved.

Recently the issue has gained renewed attention. One recommendation of the WMD Commission Report included: A structured Community program must be developed to teach rigorous tradecraft and to inculcate common standards for analysis so that, for instance, it means the same thing when two agencies say they assess something with a high degree of certainty. In addition, the 2004 Intelligence Reform and Terrorism Prevention Act states that the Director of National Intelligence shall review intelligence products to ensure, among other things, that they properly caveat and express uncertainties or confidence in analytic judgments.

Kent's proposal to use numerical probabilities with fairly wide ranges (for example, 75% plus or minus 12%) shows that there is a difference between ambiguity and lack of precision; consequently eliminating ambiguity does not require adopting a precise vocabulary. Kent's proposal was an attempt to eliminate the ambiguity inherent in the ordinary use of probabilistic terms without thereby introducing unrealistic precision. His method is one of several options that will be considered in this paper. But first the research findings will be reported.

Prior research has found that:

o Some probabilistic terms are much vaguer than others.

o People use a wide range of terms.

o People have different understandings of what the terms mean.

o Between-subject differences are larger when the terms occur in context.

o Audiences understand the terms differently from speakers.

o People understand the terms consistently over time.

Visualizing Uncertainty and Making Comparisons on Maps Using Simple Uncertainty Classes
Daniel B. Carr, (George Mason University), dcarr@gmu.edu

Abstract: Analysts in some settings have remained reluctant to provide numerical values for uncertainty. The considerations of source credibility, the absence of data and other factors can make uncertainty assessment daunting.

Given numerical values or intervals for uncertainty there are numerous possible visual encoding. These include confidence intervals about estimates, color, translucence, texture, and partitioning. Icons with labels from the National Weather Service illustrate the need for careful work in the visual encoding process. There are issues to address for each alternative representation.

This talk focuses on illustrations using dynamic shareware developed to show three variables for either arcs in network or regions on a map. The arcs can be tangible such as road or stream segments, or abstract such as in social networks. Each of the three selected variables is attached to its own two-value slider. One or more of the selected variables can be estimates of uncertainty. The sliders allow the data analyst to dynamically set the thresholds defining low, middle and high classes. One slider controls the color for a

region (or arc). The two remaining sliders define the rows and columns in a 3 x 3 grid of maps in which regions are highlighted. This design conveying for three variables and geospatial coordinates is cognitively simple and enables comparison.

The examples shown are based on publicly available data. The concepts discussed are general in nature and easily applicable to defense and national security contexts. A possible application of these methods could address the security status of road segments or regions in a city. The uncertainty of road/region security status (or length of time since a status update) and an assessment of risk can be represented and studied simultaneously. The real time response of shifting thresholds combined with capabilities such as zooming and recording of annotated views help the analyst to see and document the implications in the geospatial context.

Perhaps the reluctance to provide numerical values for uncertainty would be diminished when the interpretation is reduced to simple classes of low, middle and high and the numerical values are not tied to an absolute scale.

Wednesday, 1:30 – 3:00 Track B - **Contributed Session #1:** *"Modeling"*
Chair: Elizabeth Hohman

Performance Assessment of ISR Enterprises
Michael B. Hurley, (MIT Lincoln Laboratory), hurley@ll.mit.edu, and
Peter Jones, (MIT Lincoln Laboratory), jonep@ll.mit.edu

Abstract: The Joint Services of the Department of Defense (DoD) are in the process of transitioning legacy computing applications for multi-intelligence (multi-INT) intelligence, surveillance, and reconnaissance (ISR) tasks to standards-based information enterprises. This process is occurring with very little theoretical or practical understanding of the appropriate figures of merit needed to assess the performance of these large information systems. The danger of developing systems without clearly stated and understood figures of merit is that important, but difficult to measure, system characteristics will be ignored in favor of less important, but easier to measure, system characteristics. The result will be that the performance analyses will be at best inaccurate or at worst misleading. Operational users, system developers and program managers who rely on such performance analyses will assume considerable risk regarding their decisions on how best to use, improve and deploy these systems.

A study was conducted to develop an end-to-end assessment framework that identifies the fundamental figures of merit for multi-INT ISR enterprises to significantly reduce this risk. The study was divided into three phases: a search of the technical literature to review the state of the art for distributed enterprise frameworks similar to the ISR enterprise, the development of a conceptual framework with an analytical foundation to support performance assessment, and the construction of a simulation to demonstrate the analytical assessment of a simple multi-INT ISR enterprise.

The literature search uncovered many conceptual models and frameworks analogous to the ISR enterprise, including the Office of Force Transformations Network Centric Operations Conceptual Framework, Endsleys situation assessment model, Boyd's OODA model, and Molfetta book on network centric warfare. A few models identified quantitative metrics to assess systems, but none had an analytical foundation that was tightly integrated into the conceptual framework to define fundamental figures of merit.

To fill this need, a conceptual model with analytical foundations was developed, with the results of the literature search strongly influencing the model development. The conceptual model was developed as a sequence of three inter-related models. The first model is a black-box decision system in an environment that it attempts to modify to its advantage. The second model contains details about the black-box decision system, which consists of a sensor, a series of three information processors, and an actuator. The chain of information processors converts sensor measurements to features, then to an estimate of the state of the environment, and finally to a decision that commands the actuator to change the environment. The final model is a set of interacting decision systems that are attempting to change the world to their individual and collective advantage. The organization of the components in this final model is designed such that each

individual component adopts one of the roles in the decision system of the second-level model. This final model has sufficient detail to describe the ISR enterprise.

Analytical foundations were evaluated and selected as model development progressed from a decision system that was little more than a black box to one that is a distributed collection of decision systems. Utility theory, probability theory, and information theory were ultimately selected as the analytical foundations for the series of models. The series of models with their analytical foundation led to an important insight into the ISR enterprise: the ISR enterprise can only be fully evaluated if the sensors, command and control (C2) enterprise, actuators, and environment are included in the evaluation.

Three measures were identified by the study as being sufficient to analyze the ISR enterprise: Shannon information, the Kullback-Leibler distance, and a probability integral function. Shannon information measures the uncertainty in the information that a decision system possesses, the Kullback-Leibler distance measures the similarity between the information in two different decision systems (or components of a distributed decision system), and the probabilistic integral measures the accuracy of information that a decision system possesses when the truth is available for the evaluation. The fundamental value of these metrics is that these measures consolidate the contribution of disparate components to the ISR enterprise, including communications systems, processors, algorithms, and organizational structure, primarily by their impact on the quality of information that the enterprise can collect.

To demonstrate the application of these measures to a multi-INT ISR enterprise, a simple simulation was constructed of an enterprise that is formulating a common operational picture (COP). The simulation examined the impact that different communications architectures had on the quality of the COP by applying the decided upon metrics to the simulated enterprise. The environment for the simulation was a small five-by-five cell world with each cell either empty or containing one of three different target classes: a square, circle or triangle. The enterprise consists of five sensors with different capabilities to detect targets, ranging from a synoptic sensor that can sense the presence or absence of targets in all cells simultaneously, to myopic sensors that can only sense the absence or presence of one type of target in one cell at a time. The sensors move through the world to improve their knowledge about the world. The simulation was run with different communications architectures, including no communications, unlimited communications, push, blind pull, and informed pull. The three information measures were calculated as a function of time for multiple runs of the five communications architectures and used to quantify the performance of the different architectures. The results agree with what communications experts would predict for relative performance: the best is unlimited communication, the worst is no communication, informed pull is better than push and blind pull, despite the partial allocation of overall bandwidth to metadata transmission to support informed pull.

[This work was supported by the Department of the Navy, Office of Naval Research (ONR) under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Navy or the United States Air Force.]

Modeling the Statistical Determinants of Stochastic War Fighting Simulation Outcomes

Dr. Bill Luker Jr., (Lockheed Martin IS&S), `bill.luker@lmco.com`

Abstract: The objective of stochastic, computer-simulated combat is to draw actionable conclusions about the way changes in war fighting assets and methods can alter scenario outcomes. But in most computerized war fighting simulations, end-game output is often too voluminous or ambiguous for planners and analysts to efficiently achieve that objective. This paper discusses a set of statistical techniques borrowed from the social and managerial sciences that can help improve our ability to model whether and how new weapons systems, combinations of forces, or other changes (apart from the actual coded structure of the game itself) statistically determine simulated battlefield outcomes.

Quantitative Assessment of the National Security Strategy
LTC Michael J. Kwinn, Jr., (United States Military Academy), fm9536@usma.edu, and
MAJ Steven Gauthier, (United States Military Academy), fm9536@usma.edu

Abstract: In this paper we present a novel application of a widely practiced business application to the assessment of the National Security Strategy published by the White House in 2006. We use Value Focused Thinking (VFT) methodology to develop a hierarchical approach to assessment. This application of VFT was initially applied during Operation Enduring Freedom by the lead author as a member of a team sent to Afghanistan in August 2003.

We have taken the lessons learned from this deployment and the subsequent maturing of this approach and applied it to the National Security Strategy. After a short discussion of the approach and the process, we show the developed hierarchy and the measures for success which are taken directly from the NSS document.

This assessment methodology provides a means to indicate success or failure of the NSS over a period of time as it is a trend tool vice a picture of the state of the NSS. Though there is great reluctance to use quantitative methods for assessment of policy, this approach shows great promise in providing indicators of success or failure for the policy or policies.

Wednesday, 3:00 – 3:30 BREAK

Wednesday, 3:30 – 5:00 Track A - **Contributed Session #2:** *"Biosurveillance - I"*
Chair: Brandon Higgs

Application of a Mixed Effects Model for Biosurveillance of Regional Rail Systems
Robert J. Gallop, (West Chester University), rgallop@wcupa.edu

Abstract: Although United States government planners and others outside government had recognized the potential risk of attacks by terrorists, the events of September 11, 2001, vividly revealed the nations vulnerabilities to terrorism. Similarly, the 2004 terrorist attacks in Madrid, illustrate vulnerabilities to terrorism extend beyond the United States. Those attacks were obvious destructive acts with a primary purpose of massive causalities. Consider a bioterrorist attack which is conducted subtly through the release of a Chemical/Biological agent. If such an attack occurs through release of a specific biological agent, an awareness of the potential threat of this agent in terms of the number of infections and deaths that could occur in a community is of paramount importance in preparing the public health community to respond to this attack. An increase in Biosurveillance and novel approaches to Biosurveillance are needed. This presentation illustrates the use of mixed effects model for Biosurveillance based on commuter size for regional rail lines. With mixed effects model we can estimate for any station on a given rail system the expected daily number of commuters and establish an acceptability criterion around this expected size. If the actual commuter size is significantly smaller than the estimate, then this could be an indicator of a possible attack. We illustrate this method through an example based on the 2001 daily totals for the Port Authority Transportation Company (PATCO) rail system, which serves residents of southern New Jersey and Philadelphia region in the United States. In addition, we discuss ways to put this application in a real time setting for continuous Biosurveillance.

Characterizing bioterrorist attacks from a short time series of diagnosed patient data – A Bayesian approach

Jaideep Ray, (Sandia National Laboratories, Livermore, CA), jairay@somnet.sandia.gov, Youssef M. Marzouk, (Sandia National Laboratories, Livermore, CA), ymarzou@sandia.gov, Mark Krauss, (NORAD-NORTHCON, Colorado Springs, CO), Mark.Kraus@northcom.mil, and Petri Fast, (Lawrence Livermore National Laboratories), pfast@llnl.gov

Abstract: We present a Bayesian approach for inferring the number of infected people, the time of infection and the dosage received from an atmospheric release of an aerosolized pathogen during a bioattack. The inputs into the inference process are the number of new symptomatic patients as observed over a short (2-4 days) period, during the early epoch of the outbreak.

The release of a pathogen during a bioattack may not always be caught on environmental sensors - it may be too small, may consist of a low-quality formulation (coarse and heavy) which quickly precipitates or may occur in an uninstrumented location. In such a case, the first intimation of an attack will be the first confirmed diagnosis of a patient. Being able to infer the size of the problem from scarce data has important ramifications on the logistics of mounting a response. Further, since the estimates will be based on incomplete/incorrect observations, quantifying the uncertainty in those estimates or establishing confidence intervals becomes a concern. These estimates, once drawn, can be used in epidemic models to predict the evolution of the disease in the near future, under various levels of medical intervention. Current response plans do not contain any provisions for incorporating the uncertainty in the characterization of the outbreak at hand; they err on the side of caution by being broad and rapid. Sustainability, especially under multiple outbreaks, has not been considered an issue.

In this paper, we outline the development of the inference model [2] and apply it to a number of simulated attacks (using smallpox and anthrax) as well as the Sverdlovsk anthrax outbreak of 1979[1]. A Bayesian approach is used to develop estimates of N, the number of people infected, t, the time of infections and D, the dosage received as probability density functions, thus capturing the uncertainty in the inference. A dose-dependent incubation period model is used for anthrax [3]. Simple tests, involving people infected by an identical dosage, progress to more realistic ones where infected people receive a spectrum of dosages. This distribution is obtained by imprinting a spatially distributed population with a dosage distribution obtained from an atmospheric dispersion model. We also explore the effect of model errors i.e. where there is a systematic difference between the model used for simulating the outbreak and that used for inference. This is done by using Wilkening's Model A2 and D [3], the two models that show the closest fit to results from anthrax challenge experiments on non-human primates. Preliminary investigations [2] show that 3-5 days of data are often sufficient to arrive within a factor-of-two of the correct answer; if data is collected over 6-hour intervals rather than on a daily basis, the inferences are significantly sharper. Thus one may not require *more* data, over longer observation periods, to arrive at an accurate estimate; a better capturing of its structure, for instance through nimble reporting protocols, may be of greater assistance. Further, for diseases with long incubation periods, e.g. smallpox, the 3-5 days' observation period usually correspond to < 1% of the total infected exhibiting symptoms; however, this is usually sufficient to infer the outbreak characteristics to within very tight accuracies [2].

We finally apply this inference process to the Sverdlovsk anthrax outbreak of 1979, which, it is suspected, was caused by an accidental release of anthrax spores from a military facility [1]. 70 people died and 80 were infected. The estimated date of release is April 2nd, 1979, with the first symptoms being observed on April 4th. The symptomatic patients' time-series was reconstructed from grave-markers and interviews since much of the data had been scrubbed. Further, it is believed that the dosages were very low - estimates range from10-300 spores [3]. In addition, the progression of the outbreak had been severely modified (it lasted 42 days) by public health measures. The small size of the outbreak, the low dosages, the antibiotic-modified progression and the reconstructed data result in a stiff challenge to any inference process. Our automated method correctly identified the time of release with barely 4 days of observed data, though it took about 9 days of observations to arrive at the correct estimate for the size. Dosages were difficult to infer, though it was clear that it was less than 100 spores.

The motivation for developing this inference technique was to be able to characterize an outbreak with as little data as possible. Since this data contains noise, erroneous characterizations in the early epoch of the observation period are a constant threat. These take the form of support for hypotheses which are significantly different from the true characterization of the outbreak. We show examples of such failures and well as empirical proof that the procedure corrects itself as more data becomes available (7-8 days). While this is a measure of robustness of the procedure, length of the observation period is simply too long to be of any relevance for response planning. However, we conjecture that this particular shortcoming could be largely eliminated if prior distributions for some/all of the variables are available, since all our tests are performed with broad uniform priors. These priors are best obtained from syndromic surveillance data.

We have developed a prototypical approach for estimating the characteristics of an outbreak resulting from inhalational infection. The results discussed above encourage us to believe that such an inference process could profitably complement medical surveillance networks, by using their raw data to draw inferences regarding the size of the outbreak, including infected people still in incubation. It could also serve as a fusion mechanism for syndromic surveillance and medical reporting by exploiting priors drawn from syndromic surveillance to increase the efficiency of the inference process.

References

[1] Meselson et al, Science, 266:1202-1208, 1994.

[2] Ray et al, Sandia National Laboratories Technical Report SAND2006-1491. Unclassified, unlimited release.

[3] D. Wilkening, PNAS, 103(20):7589-7594, 2006.

Evaluation of the DC Department of Healths Syndromic Surveillance System

Arvind K. Jain, (RAND Corporation), arvind_jain@rand.org,
Beth Ann Griffin, (RAND Corporation), Beth_Ann_Griffin@rand.org,
Michael Stoto, (Georgetown University), stotom@georgetown.edu,
John Davies-Cole, (Department of Health, Washington DC), john.davies-cole@dc.gov
Chevelle Glymph, (Department of Health, Washington DC), chevelle.glymph@dc.gov,
Garret Lum, (Department of Health, Washington DC), garret.lum@dc.gov,
Gebreyesus Kidane, (Department of Health, Washington DC), gebreyesus.kidane@dc.gov,
and Samuel C. Washington, (Department of Health, Washington DC), sam.washington@dc.gov

Abstract: Immediately following September 11, 2001, the District of Columbia Department of Health (DOH) began a syndromic surveillance program based on emergency room (ER) visits. Syndromic surveillance involves the collection of data (e.g. sales volume of over-the-counter anti-nausea medications) and simultaneous analysis in order to monitor the health of a specific population. In our data, ER logs from nine hospitals are transmitted daily to the health department and categorized into mutually exclusive syndromes such as unspecified infection and gastrointestinal illness. The data are then analyzed daily using a variety of statistical detection algorithms. This paper characterizes the performance of these statistical detection algorithms in practical terms, and helps identify the optimal parameters for each algorithm given the DC DOH data as well as the most effective algorithms. Analyses were conducted to improve the sensitivity of each algorithm to detecting simulated outbreaks by fine tuning key parameters used in the algorithms. Simulation studies using the data show that over a range of simulated outbreak types, the multivariate CUSUM algorithms performed more effectively than other algorithms. Performance of the algorithms is also examined by applying them to known outbreaks such as flu seasons and a previously undetected series of gastrointestinal illness outbreaks. Our analyses appear to indicate that the DC DOH system may prove to be more valuable in identifying the beginning of the flu season than for bioterrorist attacks. The analysis also indicates that when researchers/analysts apply these algorithms to their own data, fine tuning of parameters is necessary to improve overall sensitivity.

Wednesday, 3:30 – 5:00 Track B - **Contributed Session #3:** *"Data Analysis – I"*
Chair: Nicholas Tucey

Fast and Efficient Steganalysis Methods for Spread Spectrum Steganography

Kai-Sheng Song, (Department of Statistics, Florida State University), kssong@stat.fsu.edu

Abstract: Steganography, meaning "covered writing" in Greek, is the art and science of covert/hiding communication. The rise of the Internet and multimedia techniques has prompted increasing interest in hiding information in digital media including images, audio, video, and text. The goal of steganography, like invisible inks and the microdot that were widely-used in the past century, is to convey messages under cover, concealing the very existence of information exchange, which differs from cryptography that does not conceal the communication itself but only scrambles (camouflages) the data to prevent eavesdroppers from understanding the content. Many steganographic techniques have been proposed in the engineering and computer science literature such as spread spectrum image steganography, which is a data-hiding/hidden-communication method that uses digital imagery as a cover signal. Spread spectrum steganography is analogous to spread spectrum radio transmissions where the energy of the signal is spread across a wide-frequency spectrum rather than focused on a single frequency, in an effort to make detection and jamming of the signal harder. Spread spectrum steganography takes advantage of the fact that little distortions to digital signals such as image and sound files are least detectable in the high-energy portions of the carrier (i.e., high intensity in sound files or bright colors in image files).

Many data embedding schemes in the spread spectrum steganography literature assume that the noise and interference are i.i.d. Gaussian noise and thus the schemes can be modeled as communication over a channel with, for example, additive white Gaussian noise (AWGN). However, in practical applications, the noise and interference come from signal processing and/or attack in coherent detection where the original cover signal is available, and in the case of non-coherent detection, the noise and interference consist of the host media as well as signal processing and/or attack. Consequently, the white Gaussian assumption is very unrealistic and its limitations have been demonstrated by numerous experiments. For example, different bands of block discrete cosine transform (DCT) coefficients of many natural images have shown quite different variability.

In this paper, we present a statistical formulation of the spread spectrum steganography in the frame work of generalized Gaussian families with unknown parameters of shape and possibly different variances. Both additive and multiplicative schemes of spread spectrum steganography are investigated. Stego detectors in both coherent detection and non-coherent detection are derived from the efficient scores vector. Fast algorithms for implementing the procedures are also proposed. Furthermore, the asymptotic distributions of the detectors are established. Several numerical experiments using images as a vehicle for steganographic communication are conducted to demonstrate the performance of the proposed spread spectrum steganography methods.

Achieving Human Performance for Multi-Lingual Multi-Document Summarization

John M. Conroy, (IDA Center for Computing Sciences), conroy@super.org,
Dianne P. O'Leary, (Dept. Computer Science, University of Maryland), oleary@cs.umd.edu, and
Judith D. Schlesinger, (IDA Center for Computing Sciences), judith@super.org

Abstract: Given a group of approximately 10 topically related documents in English and Arabic, compose a 100-word resume of that topic, capturing the important people, places, and details surrounding the topic event. This was the task of the 2005 and 2006 Multi-Lingual Summarization Evaluation. In this talk, I will describe a computational approach to this problem which performs at human performance levels as measured by both automatic and human evaluation.

The approach consists of three stages: a linguistic step to identify and shorten the original sentences, a statistical approach of identifying sentences with the largest expected number of terms which would appear

in a human abstract, and a linear algebraic approach for selecting a non-redundant subset of the sentences with good coverage of the important terms. See http://research.microsoft.com/~lucyv/MSE2006.htm.

Evaluation of Airborne Remote Sensing Techniques for Predicting the Distribution of Energetic Compounds on Impact Areas

Aubrey D. Magoun, (University of Louisiana at Monroe), magoun@ulm.edu,
Mark Graves, (U.S. Army Engineering Research and Development Center), Mark.R.Graves@erdc.usace.army.mil, and
Linda P. Dove, (U.S. Army Engineering Research and Development Center), Linda.P.Dove@erdc.usace.army.mil

Abstract: The characterization of impact area munitions constituents has typically employed traditional soil sampling approaches, such as stratified random techniques. These sampling approaches do not accurately account for the distribution of such contaminants over the landscape due to the distributed nature of explosive compound sources throughout impact areas, the highly localized distribution of contaminants surrounding these sources, and inaccurate records of historical target locations.

The purpose of this research was to utilize remote sensing and GIS technologies to assist in the development of enhanced sampling strategies for better predicting the landscape-scale distribution of energetic compounds and, if possible, to develop a predictive model defining contaminant source terms. Remotely sensed magnetometer and electromagnetic (EM) data, were used to thoroughly characterize metal content over a large impact area at Fort Ord, CA. This paper discusses the approaches used to develop an algorithm to better predict the landscape-scale distribution of these energetic compounds.

Wednesday, 6:00 – 8:00 Poster Session

Inventory Sampling Activity Management System (SAMS)

Donglin Sun, (U.S. Coast Guard), donglin.sun@uscg.mil,
David Hartley, (Calibre), Dave.Hartley@calibresys.com, and
Timothy Heitsch, (U.S. Coast Guard), timothy.j.heitsch@uscg.mil

Abstract: U.S. Coast Guard used the U.S. Army's Statistical Sampling System (SSS) for its quarterly audit of $1.2 billion inventory assets for over eight (8) years. Author at the Aircraft Repair and Supply Center of U.S. Coast Guard has developed Inventory Sampling Activity Management System (SAMS) since later 2003 and brought a new revolution in the history of Coast Guard Audit mandated by Department of Homeland Security. The new system developed in SAS has been running successfully for over three years and tremendously reduces previous semi-manual one-day process in SSS into 40-second event by totally automated statistical sampling process and virtually non-stopping Coast Guard inventory system to complete audit for huge inventory. Its over 500~700 times faster than SSS. The new sampling system also completes a wide range of complex sampling and reporting tasks.

CONNECTING THE DOTS: A Systems Perspective on Threat Analysis, Detection, and Prevention

Michael Round, (Center for autoSocratic Excellence), round@rationalsys.com

Abstract: The technological era has championed mass data collection as a necessary condition for analysis. For sure, great strides have been made with the backdrop of data as the foundation for a robust decision support strategy. But an interesting dilemma arises. What data do we collect to analyze? How can we be sure the data on hand is not only necessary but additionally sufficient to support reasoned analysis? What constitutes such analysis? How can we be sure?

This poster will examine the investigate techniques of Sherlock Holmes in using data on hand, targeted investigation methods, and the visual organization of results to better operationalize the connecting of dots in investigating past incidence and preventing future security threats.

Binary Synthesis Translation (BST)

Barton Clark, (Naval Surface Warfare Center Q21), barton.clark@navy.mil, and Jeffrey Solka, (Naval Surface Warfare Center Q20), jeffrey.solka@navy.mil

Abstract: Binary Synthesis Translation (BST) is a system that allows us to safely extract a portion(s) of a legacy/COTS executable. Once this code has been extracted we can then apply advanced optimization/translation operations. This approach can provide for execution of the extracted code upon a secure platform/environment. The outline of this abstract is as follows I will first discuss the work that is currently being conducted at NSWCDD with regard to binary synthesis. I will then discuss a pilot program that is currently under way utilizing our technology to mitigate critical anti tamper technology gaps that exists among several naval weapon systems that are currently being sold through foreign military sales. Decompilers have been around for a long time indeed the first known code decompiler was developed by the "Father of decompilation" Maury Halstead. In 1960, Maury directed a project on decompilation mainly to show the usefulness of the Neliac language as a Universal Computer Oriented Language, as well as a problem-oriented language. Joel Donnelly and Herman Englander implemented the D-Neliac decompiler for the Univac M-460 Countess Computer while on Maury's staff. ,Each time Joel would bring an 'impossibility' to Maury, they would sit down and figure a way around it., As Herm noted, ,the difficult we do immediately -- the impossible takes a little longer., The D-Neliac decompiler was an operational decompiler that decompiled Univac M-460 binary code into Neliac source code.

System platform dependency and application programming interface (API) routines native to a given operating system are bound into each binary image by the complier. API routines are either written in the language the compiler was written in or in lower assembler. The result of this operation is a binary program that contains not only the routines written by the programmer, but a great number of other routines linked in by the linker. A typical binary program written in C to display ``hello world'' and compiled on a COTS platform has over 22 different API subroutines in the binary program. The same Program written in Pascal and compiled on a COTS platform might generate more than 37 subroutines in the executable program. Conceptually, a decompiler works very much the same way a compiler works. It takes instructions from one format and translates them to another. The decompiler is broken up into several steps or phases.

We employ and combine several different modern decompilation techniques in pursuit of a secure synthesis solution. These techniques are based on compiler and optimization theory, and are applied to the process of de-compilation in a novel way. Our work utilizes several different components of existing Open-Source decompilers. They include dcc a UNIX based command line decompiler written by Cristina Cifuentes. Boomerang is a UNIX/Windows based GUI system that is probably the most mature of the major decompilers in use today. Boomerang is primarily developed by QuantumG and Mike Van Emmerik. Andromeda is a GUI version for Windows developed by Andrey Shulga. Our BST can be characterized as a system composed of several phases which are grouped into modules dependent on language or machine features. The front-end is a machine dependent module that parses the binary program, analyzes the semantics of the instructions in the program, and generates an intermediate representation of the program. We generate a control flow graph of each subroutine. BST can operate with language and machine independent modules. The system analyzes the low-level intermediate code and transforms it into a high-level representation available in any high-level language, and analyzes the structure of the control flow graph(s) and transforms them into graphs that make use of high-level control structures. Finally, the back-end is a target language dependent module that generates code for the target language.

BST benefits from compiler and Application Programming Interface (API) signatures resident within a binary image. In the former compiler signatures for any start-up code is ignored and not decompiled. In the latter any API references are used for variable type information and propagated thru out the function analysis process. The BCS system is comprised of the three following modules each with a set of corresponding sub-modules:
Front-end: (Machine Code Dependent)
Syntax Analyzer
Semantic Analysis
Intermediate Code Generation

Control Flow Graph Generation
Analysis (machine code Independent)
Global Data Flow Analysis
Back end
Code Generation, Optimization, Translation
Syntax Analyzer

The syntax parser analyzer groups bytes of the source program into grammatical phrases (or sentences) of the source machine language. These phrases or Idioms are stored in a hierarchical tree. One limitation with machine code is that the hierarchy will always have a maximum of two levels. The main problem encountered while parsing machine code is determining what data is and what an instruction is. For example, a case table which usually resides after the function that invokes it will be located in the code segment and is unknown to the decompiler weather the table is data or instruction. This is a common problem with COTS memory architectures that utilize the von Neumann architecture (Data and Code reside in same memory). One can not simply make the assumption that instructions can be parsed sequentially assuming that the next byte will always hold an instruction. Many machine dependent heuristics are needed in order to determine the correct set of instructions Semantic Analysis This phase involves checking the source program for the semantic meaning of groups of instructions. We gather the type information, and propagate this across the subroutine. If we safely assume that binary programs were produced by a compiler, the semantics of the machine language is correct in order for the program to execute (assuming the program executed properly). We make the assumption that semantic errors will not be present in the source program unless the syntax parser has performed an error such as data has been parsed instead of instructions Intermediate Code Generation We need to generate an intermediate representation of the source program so that the decompiler can analyze low-level structures within the module. During the generation process a road map is followed that will allow easier migration to the target language desired.. During this phase we utilize the three-address code instruction mapping.

Control Flow Graph Generation We generate a control flow graph of each subroutine in the source. This approach can be very useful for removing dead code, obfuscations, determining def-use relationships across procedures etc. We can additionally determine highlevel control structures used in the program. This graph will also assist in removal of compiler generated intermediate jumps Data Flow Analysis During this phase we attempt to refine the intermediate code. Here we begin to identify high-level language expressions. We try to remove any temporary registers or condition flags as these constructs are not used in high-level languages. This process involves determining data-dependencies and defined usages within a basic block of code. We use API signature type information to assist in identification of variable types. This information is then used thru out the procedural data scoping process. Once the procedure data scope has been established, then a higher order inter-procedural data scope is obtained. During this process complex global data flow analysis equations are solved for each procedure. This includes any parameters that are referenced by the procedure and any return values, and any global data variables are modified inside the procedures.

Code Generation During the final phase or back-end of the process, we begin the process of producing target source-code. The target language needs to be specified along with any optimizing or specific in-lining options. Traversal of the control graph for each sub-routine is implemented to handle such issues as variable naming, local stacks arguments, and register variable identifiers . Additionally the control-structures and intermediate instructions that were created in earlier steps are now translated to high-level language statements.

The discussion above is primarily concerned with converting machine language to a higher-level language source. The particular weapon system that we are protecting uses machine language that was originally derived from assembly language source. This system never used a high-level language like ,C, or FORTRAN.

Our approach with the SPY 1-D radar combat system is to identify CPI code to extract. We then disassemble the machine binary into an internal representation. We then generate control flow graphs, conduct semantic analysis ,data flow analysis this includes single static assignment (SSA), data propagation,

register variable identification, data type propagation, data type analysis, primitive data types, complex data types, control flow analysis. We used these techniques to successfully extract a function or functions void of any data dependencies and API dependencies. This renders a sequence of code that will execute outside its normal environment. It is important to note that after CPI extraction, we must patch the host executable where CPI code once resided. If for example the host executable contains one CPI function that needs to be extracted, then we must patch the function entry point of that CPI function. We need to be able to invoke the external CPI function from the host machine. We handle this problem by replacing the next instruction after the function entry point with an interrupt service routine (ISR) call. This ISR routine will reside within the host ISR Table. The ISR routine can examine the contents of the stack owned by the process that invoked the ISR. We examine the stack for a return address. This return address will point back to the location directly following the ISR invocation. This information will allow us to determine which CPI function is invoking the ISR routine. One issue remains for the CPI to execute remotely. If for example the CPI function calls an operating system API call that in turn issues an interrupt service routine (ISR) request, then it is necessary to provide a mechanism that allows for the CPI function to execute the ISR instruction (on the host) remotely.

We mitigate this problem by providing on the host (target) machine and on the external secure board a remote procedure call interface (RPC) module that includes both client and server capability. This for example allows a CPI function to issue an ISR call while executing remotely.. In the case of the CPI function issuing an ISR call back to the host machine the CPI call would be issued within the context of a remote procedure call client. The host machine would be functioning within the context of a remote procedure call server. Once we have completed transforming and extracting the CPI code from the original executable we then package a remote relocatable library. This library module can be initially tested with in the original host executable process environment. We perform this operation to ensure that all dependency (memory, system call) operations have been resolved correctly. This process is very similar to executing a program and then at run time issuing a system load library call to load a given dynamic library into a process space. We conduct a validation and verification operation to insure proper operation of the CPI. The CPI code is then written out to a secure Field Programmable Gate Array (FPGA) chip residing on an external board. The FPGA employs a 128 bit Non-Volatile encryption strategy. This key can not be extracted externally. At this point we can determine what type of core execution image will reside within the (FPGA). If we decide to use a soft core execution environment (EE) the system will run at a greatly reduced speed but with greater flexibility. If we use a hard-core EE image the system can execute up to 800 MHz. If we decide to change the EE of the FPGA it will be necessary to translate the CPI code to match the EE of the FPGA.

Naturally this approach introduces a degree of operational latency by virtue of passing thru the ISR and RPC code. This is of course not a problem with the 1750 CPU found in the SPY-1D which currently runs at 33 MHz. The only real timing problem in this case is not to return too early back to the host process. In the case of newer COTS systems however while typical CPU's are executing internally at 3.0 GHz. range they are still only able to fetch instructions and memory at local bus speeds which are typically in the 800MHz. range.

The Altera FPGA we are using operates in the 800 MHz. ranges. Since we are using BST to rebuild the extracted function(s) we an employ such techniques as Function-Level Working-Set tuning. We can profile functions that are executed with in a target executable. The functions that are executed more frequently than others can be moved closer to the top of the module. This way the operating system can keep the popular code in memory and only load the rest of the module as needed (and then page it out again when it's no longer needed). This approach can provide for a significant increase in speed as it reduces on demand memory paging. We can further increase the speed of the extracted function by implementing strategies like reciprocal multiplication. The idea with reciprocal multiplication is to use multiplication instead of division in order to implement the division operation. Typically multiplication is four to six times faster than native division operations. The idea is to multiply a dividend by a fraction that is the reciprocal of the divisor. For example, if you wanted to divide 30 by 3, you would simply compute the reciprocal of 3, which is one divided by 3. The results of such an operation is approximately 0.3333333, so if you simply multiplying 30 by 0.3333333, you'd end up with it the correct result, which is 10.

We can further optimize by deconstructing an instruction and implementing its micro-code underpinnings. If we examine a typical floating point instruction found in many CPI algorithms. We can see that much of these instructions are implemented as lower-level multiply and divide operations. We can apply symmetric parallel optimization during the construction process and we can fabricate a custom micro-instruction that will utilize symmetric parallel optimization to run much faster than the original instruction.

The FPGA includes additional I/O pin tamper detection safety circuitry that is crucial to prevent any type of black-box attack. If an individual were to gain possession of the FPGA and probe with arbitrary input signals, the Altera provides a tamper detection capability. It works as follows: Learning Phase, the Input Monitoring Model is trained from real or simulated inputs and outputs to the device to be protected. These inputs and outputs could be measured, obtained from the device specification, or extrapolated from piecemeal knowledge of the device's characteristics. Once this database of known input-output combinations (both normal operational inputs and simulated tamper-style inputs) is created, it will be applied to a time-sensitive Input Monitoring Model.

Operational Phase, the component under protection is either considered to be in its normal operational environment or it is outside of this environment and being subjected to laboratory attacks. External inputs are received by the augmented anti-tamper package and fed directly to the protected component, the Input Monitoring Model and the Output Obfuscation Model. These inputs come from either the intended environment or from laboratory testing; the component is unaware of the source initially. In the former case, the Input Monitoring Model will respond with a No Tamper signal. This signal is used by the Output Obfuscation Model's Gating Mechanism to pass the normal operational outputs from the critical component through to the output of the augmented package. If, however, the recognizer outputs a Tampered Input signal, this signal initiates processing within the Output Obfuscation Model. Based upon the temporal inputs being received and the indicator from the recognizer, the generator will produce obfuscated outputs that are then passed through the gating mechanism to the output.

BST holds promise to solve many complex software vulnerability problems present in today's modern computer systems as well as older legacy systems. BST can help identify polymorphic viral code that falls under the radar screen of today's current signature based anti-virus products. BST can be used to mitigate buffer overflow vulnerabilities in legacy/COTS code by extracting code that is unsecured and running it within a secure framework. BST can be used to prevent Reverse Engineering (RE) efforts by protecting CPI code in a secure shielded environment particularly software vendor's protection mechanisms can be extracted and executed in a secure environment as well as military weapon systems CPI content. BST can be used to assist in software optimization and or upgrading of obsolete legacy code to modern software libraries.

Exploratory Data Analysis on Document Collections

Nicholas Tucey, (Naval Surface Warfare Center Dahlgren Division), `nicholas.tucey@navy.mil`

Abstract: A demonstration on various text data mining software tools will be given. The software indexes a collection of documents using the Java Lucene search engine API and Latent Semantic Indexing (LSI). Both indexing techniques allow the user to explore the document collection using queries. Visualization of the collection is performed using spectral graph projections and hierarchical clustering of the documents. Finally, a social network visualization tool will be demonstrated using the collaboration of authors from the document collection.

WEBSTER AGENT CASE EXPERT (ACE) DEMO

Adolf Neumann, (21st Century Systems, Inc.), `adolf.neumann@21csi.com`

Abstract: The WEBSTER Program's Agent Case Expert (ACE) software provides distributed collaborative trust networks and opinion summary mechanisms for intelligently integrating decisions made by both man and machine. ACE explicitly accounts for dynamic trust and pedigree in a semantically-tagged decision model using extensible belief algebra for deriving semantically coherent consensus. ACE trust network can model evidential opinion flow through social and organizational networks. ACE trust network

derivation is guided by a semantic decomposition according to subject domain knowledge. The ACE trust mechanism allows for information to be reused and integrated despite changes in trustworthiness due to factors such as information aging or relevance. ACE provides a normalization layer for integrating heterogeneous, both human and mechanistic, intelligence analyzers, characterizers, and classifiers. This presentation will provide an interactive demonstration of this process, and permit the viewers to ask in-depth questions of one of the key architects of the WEBSTER ACE software.

Assessment of Unmanned Ground Vehicle Technology in an Operational Context

Barry Bodt, (US Army Research Laboratory), babodt@arl.army.mil, and
Marshal Childers, (US Army Research Laboratory), marshal.childers@us.army.mil

Abstract: Robotics is an important area of research in the defense and national security arena and is worthy of this conferences attention. Unmanned air, ground, sea, and subsurface systems are used or proposed as force multipliers for any number of military tasks, for example, logistics, force protection, reconnaissance, and assault. Whole organizations, such as the Association of Unmanned Vehicle Systems International are dedicated to its pursuit.

The U.S. Army Research Laboratory (ARL) manages both basic and applied research programs focused on advanced perception, intelligence, and human-robot interaction. While the program was initially focused solely on autonomous mobility, the success achieved to date in autonomous navigation has led the Army to increase its focus to issues more closely associated with fielded systems, e.g., operator workload.

To examine the impact of technology upon operational issues, an experimental unmanned vehicle (XUV) was equipped with a live reconnaissance, surveillance, and target acquisition (RSTA) capability. Where live targets were not readily available, target chips were simulated for random presentation. Beyond the RSTA technology, there are operator workload considerations. Although autonomous mobility has been achieved to a degree, maintenance of autonomous operations is still required when the XUV encounters certain situations. This remote handling takes time and competes with the time required to complete tasks for a RSTA mission. As the operator performs tasks in both work areas, man-machine interface must be assessed.

In this poster, we will provide an overview of what has been tried from an experiment and measurement standpoint to assess unmanned ground vehicle use in one operational context. Another round of experimentation is scheduled for 2007 to assess further advancements. We welcome any insight or suggestions the conference might offer in the areas of (1) workload assessment, (2) experimental design, or (3) measures of performance with regard to this program.

Thursday, 8 February 2007

8:30 - 10:00 Track C - **Topic Contributed Session#1:** *"Data Fusion"*
Organizer: Ed Wright
Chair: Wendy Martinez

Multi-Semantic Fusion
Ed Wright, (IET, Inc.), ewright@iet.com, and
Masami Takikawa, (IET, Inc.), takikawa@iet.com

Abstract: In traditional data fusion applications, fusion is performed using data from sensors that are reasonably well understood and provide error models with known parameters. In today's operational asymmetric warfare environment, however, it is required to fuse information from a much broader set of sources (e.g, HUMINT, COMINT intercepts, and open source). Fusing information from such diverse sources necessitates handling of additional uncertainties about the meaning or semantics of data and characterization of sources. Such semantic fusion poses various challenges, including interpreting information (e.g., texts in various natural languages or outputs from sensors not well understood), mapping and aligning different ontologies, inferring credibility of sources, and integrating information in various forms of uncertainties such as probabilities, Dempster Shaffer (D-S) belief functions, and fuzzy logic.

We use Bayesian Networks as the common representation for alternative representations of uncertainty, including D-S and fuzzy-logic, in order to fuse information from multiple sources. We have developed probabilistic representations for use in hierarchical Bayesian inference, and the methodology for converting alternate uncertainty representations to probabilities. The development includes the methodology for learning probabilistic models of legacy systems when the existing representation is unknown. Our approach requires that we be able to incorporate D-S, fuzzy systems, and other ad hoc or unknown uncertainty representations.

For D-S we use the plausibility transformation from D-S to BN to obtain probabilities by using the plausibility function computed from the D-S belief function. Unlike other transformations (e.g., the pignistic one), this is compatible with Dempster's rule, and exactly the same results can be obtained by using efficient Bayesian inference.

For systems that use fuzzy membership to represent uncertainty, we represent the fuzzy report as evidence from a source with an unknown credibility model. If details of the fuzzy membership functions are available, that information is used to instantiate the credibility model. When no details are available, we recognize that there are many transformation methods available among fuzzy-logic membership functions, possibilities, D-S belief functions, Random-sets, and probabilities. We include second order uncertainty in the credibility model to represent the range of possible transformations. When multiple observations are available from a consistent fuzzy source, the BN model allows us to learn the parameters of an appropriate credibility model.

Systems with ad hoc or otherwise unknown uncertainty representations are also modeled using a generic credibility model with second order uncertainty. When multiple observations are available from a consistent fuzzy source, the BN model allows us to learn the parameters of an appropriate credibility model.

In addition to the theoretical and modeling work, we have created a simple model of fuzzy sensors and shown how a second-order uncertainty model can be used to interpret fuzzy sensor reports. We have also shown how these models allow us to change the interpretation of existing reports upon receiving new information about sensors or multiple reports from the same sensor.

Bayesian networks are also used to implement a computational model that characterizes the semantics of information. Such a computational model can be used to (1) incorporate semantic information in the fusion process; (2) deal with missing or uncertain semantic information; and (3) update results of previous inference to make use of new semantic information when it becomes available.

Benefits of using Bayesian networks as a semantic fusion framework include the ability of making coherent and optimal decisions using well-established Bayesian decision theory, a compact representation which makes knowledge elicitation and inference more tractable, the availability of advanced BN representation and inference mechanisms such as Multi-Entity Bayesian Networks (MEBNs) which combine Bayesian networks with first-order expressibility, enabling the creation of modular and reusable probability models.

In this paper, we describe MEBN models for semantic fusion with various uncertainty representations (e.g., fuzzy systems), and demonstrate their ability to fuse information in various uncertainty representations, make inference about the characteristics of unknown sources, and update fusion results with new semantic information.

Inducing Observables to Detect Hidden Files and Processes on Computer Systems
Jim Jones, (SAIC and Ferris State University), jonesjame@saic.com

Abstract: File and process hiding are standard features of rootkit applications which are used to maintain computer systems in a compromised state. Such features have recently appeared in malicious applications like viruses, worms, and spyware, as well as legitimate applications such as digital rights management. File and process hiding methods and implementations continue to evolve in order to evade detection, and current detection approaches are ineffective against these latest methods and implementations. Computer systems which are in a compromised state but believed to be clean due to scans using current methods represent a considerable risk to computer users and owners. We report on an approach to detect hidden files and processes on potentially compromised computer systems. We specifically target detection of files and processes which have been hidden using novel methods against which current approaches are ineffective. Our approach consists of probes to induce observables and a probabilistic model to reason over the induced observables. We discuss the development of the probes and the probabilistic model, and we report on empirical comparisons of our approach to current detection methods.

Bayesian Ontologies in Net-Centric Systems
Kathryn Blackmond Laskey, (George Mason University), klaskey@gmu.edu, and
Paulo C.G. Costa, (George Mason University), pcosta@gmu.edu

Abstract: Ontologies have become ubiquitous in current-generation information systems, and are a key enabling technology for Net-Centric Warfare. An ontology is an explicit, formal representation of the entities and relationships that can exist in a domain of application. Following a well-trodden path, initial research in computational ontology has neglected uncertainty, developing almost exclusively within the framework of classical logic. As appreciation grows of the limitations of ontology formalisms that cannot represent uncertainty, the demand from user communities increases for ontology formalisms with the power to express uncertainty. Support for uncertainty is essential for interoperability, knowledge sharing, and knowledge reuse. Bayesian ontologies are used to describe knowledge about a domain with its associated uncertainty in a principled, structured, sharable, and machine-understandable way. This paper considers Multi-Entity Bayesian Networks (MEBN) as a logical basis for Bayesian ontologies, and describes PR-OWL, a MEBN-based probabilistic extension to the ontology language OWL. To illustrate the potentialities of Bayesian probabilistic ontologies in the development of AI systems, we present a case study in information security, in which ontology development played a key role.

Thursday, 8:30 - 10:00 Track D - **Topic Contributed Session #2:** *"Text Data Mining"*
Organizer and Chair: Jeffrey Solka

Discovery Facilitation Via Latent Semantic Indexing

Jeff Solka, (Naval Surface Warfare Center), jeffrey.solka@navy.mil,
Nicholas Tucey, (Naval Surface Warfare Center), nicholas.tucey@navy.mil, and
Avory Bryant, (Naval Surface Warfare Center), avory.bryant@navy.mil

Abstract: Previous work by Swanson and other (Swanson, 1986, Smalheiser, NR & Swanson, DR., 1998, Gordon, M.D. and Lindsay, R. K., 1996) have developed methodologies for the semi-automated detection of new discoveries. These new discoveries have consisted of new candidate approaches to standing problems. For example Swanson himself had applied his developed methodologies to discovery of new techniques for the treatment of Raynaud's syndrome (Swanson, 1986), migraines (Swanson, D. R., 1988), and even the prevention of technological surprise (Swanson D.R, Smalheiser N.R and Bookstein A., 2001).

Swanson's original approach sought out transitive links between related literatures. Gordon and Lindsay 1996 sought to automate Swanson's procedures through the application of statistical down selection procedures. Our work has followed the lead of the previous efforts of Gordon and Dumais (Gordon, M. D., and Dumais, S., 1998) that utilizes latent semantic indexing (LSI) to facilitate the discovery process. Their work primarily focused on the identification of related terms.

The reader is reminded that LSI is based on singular value decomposition on the term document matrix and that this decomposition provides projections that allow one to render the terms and documents within a common space. Gordon and Dumais used the projection that rendered the terms (single and double terms) to look for interesting associations between the terms that were originally associated with the problem of interest and those that might offer potential solutions. They discussed within the venues of their paper the fact that one could take a similar approach to look for associations among documents rather than terms.

This talk will discuss some of our recent work to revisit and extend the methodology of Gordon and Dumais. Our work has focused on the development of visualization frameworks and GUI-based software systems to facilitate the identification of potential discoveries based on term to term and document to document associations. We will illustrate these approaches on a current problem of interest which is the discovery of new methods of water purification.

References

Gordon, M. D., and Dumais, S., 1998. Using latent semantic indexing for literature-based discovery. Journal of the American Society for Information Science. 49(8): 674-685.

Gordon, M.D. and Lindsay, R. K., 1996. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science. 47. 116-128.

Smalheiser, NR & Swanson, DR., 1998. Using Arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57, 149-153.

Swanson, D. R., 1986. Fish Oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1), 7-18.

Swanson, D. R., 1988. Migraine and magnesium: eleven neglected connections. Perspt. Biol. Med., 31, 526-557.

Swanson D.R, Smalheiser N.R and Bookstein A., 2001. Information discovery from complementary literatures: categorizing viruses as potential weapons. JASIST 52(10), 797-812.

Measuring Word Frequencies for an Evolving Lexicon
Elizabeth Leeds Hohman, (Naval Surface Warfare Center), elizabeth.hohman@navy.mil

Abstract: This work is part of a larger project to analyze streaming documents such as news articles or web logs. The project uses a graph representation of the documents and provides dynamic methods for clustering and viewing the documents. As part of that project, a vector space model is used to represent the documents. In the vector space model, each document is represented as a vector with each dimension of the vector corresponding to a different word in the lexicon. The entries of the vector depend on the number of times the corresponding word in the lexicon occurred in the document. The entries are usually scaled by the frequency of the word in the corpus. This decreases the effect of common words that occur in many documents and increases the effect of rare words that signify the content of the document.

The data in this project are considered in a streaming fashion such as news articles or newsgroup entries collected in time. If we no longer assume a fixed corpus, we cannot use a fixed-dimensional vector space model. Since the lexicon cannot grow without limit, approximations must be made to the representation. New words will appear in documents while words that have been seen in the past might not be seen again. Since the lexicon is constantly changing, we cannot pre-assign dimensions of the vector space to specific words. Also, since vector entries are dependent not only on the frequency of the word in the document but also on the frequency of the word in the corpus, the corpus frequency must also be approximated in the case of streaming documents.

One solution to approximating the corpus frequency is to use a time window and calculate the frequency within that window. In this work, we use an exponentially weighted moving average where the parameter allows for varying the amount of history influencing the value. Since the parameter influences the amount of emphasis on past documents, we should expect streaming text sources that are changing more rapidly to require a different value than sources which change at a slower rate. We will examine simple classifiers and simple datasets in order to monitor classification performance as a function of this parameter. Although the focus will be on calculating word frequencies for the evolving lexicon, other details of text processing of streaming documents will also be presented.

Understanding Logistics Through Text Analysis
John Rigsby, (Naval Surface Warfare Center), john.rigsby@navy.mil, and
Dr. Jeffrey Solka, (Naval Surface Warfare Center), jeffrey.solka@navy.mil

Abstract: This paper will present the application of text data mining (TDM) methodologies to obtain accelerated in-depth understanding of the state of the art in the field of transportation and distribution logistics.

Logistics document collections were developed from Defense Technical Information Center (DTIC) technical reports. The purpose of the analysis is to create and analyze a compilation of literature to understand the vocabulary, sub-topics, and domain knowledge encompassing the field of logistics.

Thursday, 10:00 – 10:30 BREAK

Thursday, 10:30 – 12:00 Track C - **Topic Contributed Session #3:** *"Network Analysis"*
Organizer: David Marchette
Chair: John Rigsby

Predicting Unobserved Links in Covert Networks
David Marchette, (Naval Surface Warfare Center), dmarchette@gmail.com

Abstract: In covert networks it is often the case that some links will be unobserved: it will be unknown whether potential edges are present in the graph or not. I will discuss a simple generative model for social networks, the constrained random dot product graph (CRDPG), and illustrate how this can be used to suggest high probability edges for further investigation. The model will be illustrated on a dataset

consisting of alliances between nations. These data have covariates on the vertices which will be used to improve the performance of the prediction. Other datasets will be discussed as time permits.

Interaction Models of Probabilistic Networks Suggested by Statistical Mechanics

John E. Gray, (Naval Surface Warfare Center), john.e.gray@navy.mil

Abstract: Statistical Mechanics has proven to be useful model for drawing inferences about the collective behavior of individual objects that interact according to a known force law (which for more general usage is referred to as interacting units.). Collective behavior is determined not by computing F=ma for each interacting unit because the problem is mathematically intractable. Instead, one computes the partition function for the collection of interacting units and predicts statistical behavior from the partition function. Statistical mechanics was unified with Bayesian inference by Jaynes who demonstrated that the partition function assignment of probabilities via the interaction Hamiltonian is the solution to a Bayesian assignment of probabilities based on the maximum entropy method with known means and standard deviations. Once this technique has been applied to a variety of problems and obtained a solution, one can, of course, solve the inverse problem of to determine the solution to an inverse problem to determine what interaction model gives rise to a given probability assignment. Probabilistic networks are important modeling tools in a variety of applications including social networks. I explore the usage of statistical mechanics as a mechanism to solve the inverse problem for a probabilistic network to determine what the underlying interaction model that gives rise to probabilistic network. Then I explore how one can draw general inferences about the network by defining, energy, heat capacity, temperature and other thermodynamic characteristics of the network.

Using Social Network Analysis to Evaluate Public Health Preparedness

Leslie McIntosh, (Saint Louis University), mcintold@slu.edu

Abstract: For this research, I propose using social network analysis (SNA) as a viable means to gauge the preparedness level of communities who participate in multidisciplinary emergency response exercises. With millions of dollars being spent on public health preparedness programs and trainings, there are few methods by which to evaluate these programs. Further difficulties exist due to the many facets of public health preparedness which need to be assessed from individuals competencies and skill sets to organizational capacity and participation levels.

In a three-part evaluation, the awareness, performance, and integration abilities of participants are assessed. Awareness is the measurement of who knows what information; while performance is the measurement of who has what skills. Participants are rated as proficient or not proficient in their awareness and performance. The integration aspect is characterized through multiple measurements (i.e. density, centrality) using SNA and depicts how the participants relate to one another during the exercise.

Awareness proficiency is assessed using existing tools such as the National Incident Management System (NIMS) competencies, and performance proficiency is measured using metrics established from sources such as Target Capabilities List (TCL).

A social network is formed that retains information of individual awareness and performance abilities while assessing the group-level integration capacity. From this information, I look at the shape and density of the network, along with measuring the ability to disable the network. The shape of the network can take on many forms such as hierarchical or cellular, while the density of the network ranges from 0 (no participants are connected) to 1 (everyone is connected). In addition, I hypothesize that the strength of the network can also be revealed using statistical algorithms that disable the network.

Together, these data and analyses elucidate community levels of preparedness. This information can then be used to visually and statistically depict emergency response exercise preparedness. These data then become the base-line measurements that can be used for follow-up evaluations of future exercises.

Thursday, 10:30 – 12:00 Track D - **Topic Contributed Session #4:** *"Operations Research"*
Organizer and Chair: Don Wagner

Hierarchical High Level Information Fusion Using Graph Structures, Subgraph Matching and State Space Search

Moises Sudit, (University at Buffalo), sudit@eng.buffalo.edu,
Rakesh Nagi, (University at Buffalo), nagi@buffalo.edu, and
Kedar Sambhoos, (University at Buffalo), kps6@buffalo.edu

Abstract: An enormous amount of research has been conducted in the area of multi-sensor data fusion. Under the Joint Directors of Laboratories (JDL) five-level data fusion model, Level 1 on Object refinement seems to have received the most attention. Level 1 processing functions include: data alignment, association, tracking, and identification. Less mature are Level 2 processing, situation assessment, which seeks a higher level of inference above level 1 processing, and Level 3 processing (for military or intelligence fusion systems) which performs threat assessment. The purpose of threat assessment is to determine the inherent threat of a situational state estimate using an inferential process. Level 2 processing (Situation Assessment) attempts to create an understanding of the knowledge of objects, their characteristics, relationships among objects and cross force relations. Situational assessment is the process of providing understanding to a decision-maker about a situation, including people, objects, events, and the environment. At the lowest level, assessment systems track information about relationships, movement, and classification. More sophisticated systems include understanding of the context, estimation of intent, and, ultimately, adaptation of the system to its particular environment.

The intent of this presentation is to show enhancements in Level 2 and 3 fusion capabilities through a new class of models and algorithms using graph structures. The problem today is not often lack of data, but instead, lack of information and data overload. Graph matching algorithms helps us solve this problem by identifying meaningful patterns in volumous amounts of data to provide information. In this paper we investigate a classical graph matching technique of subgraph isomorphism. A complete implementation of a heuristic approach (since the problem under consideration is NP-Hard) using an inexact isomorphism technique has been used. The heuristic approach is called Truncated Search Tree algorithm (TruST), where state space of the problem is constrained using breath and depth control parameters. The breath and depth control parameters are then studied using design of experiment based inferential statistics. Finally, a software implementation of the procedure has been completed with very encouraging empirical results.

Graph Matching is a classical optimization approach that has been studied for a number of years and for which at least two independent graphs are required. The first graph is called a Data Graph (also Input Graph), which contains all the sensor information gathered on a specific domain of interest. The second graph is called a Template Graph (also Template or Pattern), which describes a predefined information signature which is of relevance to an analyst. In general the Template Graph is much smaller than the Data Graph and the objective is to investigate the syntactic and semantic occurrence of Template Graph in the DataGraph.

The primary objective of this work is the progression of Level 2/3 fusion of informational content to obtain an advanced multi-intelligent system for hierarchical high level decision making processes. The goal of the proposed work is to develop an information integration mechanism to simplify human decision making solving operational problems. As technology continues to advance, and the proliferation of sensors in all platform increases, human decision makers are being overwhelmed with data. In summary, we use attributed graph models to represent situations in Level 2 and 3 Fusion. Graph matching is invoked to determine if a situation of interest to the analyst exists in a scenario. A Truncated Search Tree heuristic is developed to perform graph matching. A Maritime Domain example will be shown where hierarchical information fusion will be tested.

Multi-threat Containment with Cooperative Autonomous Agents
Shanchieh Jay Yang, (Department of Computer Engineering, Rochester Institute of Technology), jay.yang@rit.edu , and
Bhushan Mehendale, (Rochester Institute of Technology), bhushan@bhushan.in

Abstract: The study of sensors and robotics has move beyond optimizing individual system performance. A key focus today is on overcoming challenges of utilizing a large number of cost-effective, autonomous, cooperative agents (sensors or robots). This work will lay out a few key open problems in the field, with a focus on the multi-threat containment problem. The multi-threat containment problem asks a set of autonomous agents to engulf and observe occurring threats aided by only local sensing capabilities. Previous related work has dealt with robot formations and single threat containment, in which cases a single target exist throughout the containment process. In order to accommodate multiple threats occurring at random times and expires, autonomous robots need to dynamically adjust their target threats to ensure containments. A potential-field based approach is taken in this work to enable distributed and multiple threat containment and collision avoidance. Each agent in the field will independently form its own view (the potential field) of the surrounding; this view will be periodically updated and used to determine the agent's direction and velocity in traveling to engulf the target threats. A combination of quadratic potential functions is used to model the potential field seen by the agents, which choice simplifies the computational complexity and the parameter design. The simplicity of the proposed algorithm, Multiple Threat Containment Algorithm (MUTCA), is expected to enable its realization on cost-effective robots (~$200). MUTCA has been simulated with different threat occurrence settings as well as robot capabilities. The talk will demonstrate the simulation results, showing the benefits and the limitations of MUTCA, and conclude with a discussion on future directions of this research.

Maritime Tracking of Past Data
L. D. Servi, (MIT Lincoln Laboratory), servi@ll.mit.edu

Abstract: Maritime tracking, and maritime domain awareness more generally, has gained increasing interest in recent years as typified by a recommendation for improved systems in the Quadrennial Defense Review (QDR) published by the Department of Defense on February 6, 2006 (page 58). Maritime tracking problems can be segmented as real time or forensic depending on whether the track is estimated while the data is collected or after it is collected. This talk will concern only the latter case.

Fast forensic tracking, i.e., tracking a ship's previous path, could be useful in a number of applications in defense and national security. For example, the Coast Guard must approve all ships entering an US port and hence may be assisted by improved information about the ship's previous path. Alternatively, after a catastrophic event occurs at sea there is a need to forensically examine the ship(s) involved to assist in inferring attribution, finding the perpetrators, and/or identifying relevant perpetrator infrastructure related to the event. Finally there may be a desire to periodically update the location of suspicious but not imminently dangerous ships.

It is empirically useful to assume a ship's path consists of a concatenation of a number of paths parameterized by a small number of variables. In its simplest case, it could be idealized by a line segment or a piecewise linear paths.

This talk will describe insights into this forensic tracking problem first from the image processing literature point of view with a focus on Hough Transforms. Here the basic approach is to parameterize the space of the potential ship paths, construct a goodness of fit objective function, and then maximize the objective function. What makes this approach difficult is the largeness of the state space and the highly non-convex objective function found in realistic settings.

An alternative approach, which has been thus far been used only under the assumption of precise location measurements, will be presented and its performance illustrated using simulated data. In particular, it will be shown that the algorithm can identify a path consisting of 10 true measurements amid 1000 false measurements in .047 seconds and can identify a path consisting of 1000 true measurements amid 10,000

false measurements in .38 seconds (both on a dual 3 GHz Xeon processor). For the case of imprecise location measurement, directions of future efforts will be summarized.

*This work was sponsored by the U.S. Government under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Thursday, 12:00 – 1:30 LUNCH

Thursday, 1:30 – 3:00 Track C - **Contributed Session #4:** *"Biosurveillance - II"*
Chair: Chris Overall

Approximating the Sum of Lognormal Distributions to Enhance Models of Inhalational Anthrax

William R. Hogan, MD, MS, (Department of Biomedical Informatics, University of Pittsburgh), wrh@cbmi.pitt.edu, and
Garrick L. Wallstrom, PhD, (Department of Biomedical Informatics, University of Pittsburgh), garrick@cbmi.pitt.edu

Abstract: In many biological defense applications, it is useful to model the incubation period of disease. Numerous authors have simulated inhalational anthrax, either to analyze response policies or to evaluate outbreak detection systems [1-4]. Each of these authors used a mathematical description of the incubation period of inhalational anthrax, and at least two authors have written entire papers devoted to the topic [5, 6].

However, many authors omit from their models of inhalational anthrax the time from symptom onset to the time an individual presents for medical care. We refer to this interval as the visit delay. One possible reason for excluding visit delay is that it is expected to be much shorter than the incubation period for anthrax. A second reason may be the mathematical difficulties involved with computing the sum of the distributions used to model incubation period and visit delay.

By far the most common distribution that researchers use to model the incubation period of anthrax--and infectious diseases in general--is the lognormal distribution. The reason is that it is simple and it describes adequately the distribution of incubation periods for a wide variety of diseases [7, 8]. We therefore chose to model visit delay using the lognormal distribution as well. However, if we model incubation period and visit delay with lognormal distributions, then the distribution over the interval from exposure to presentation for medical care is the sum of two lognormal distributions, for which no closed form solution exists. This situation is not a problem for simulating individual cases of anthrax: simply generate a random variate from the lognormal distribution for the incubation period and a random variate from the lognormal distribution for visit delay and sum them to obtain the time the individual presents for medical care. However, to compute efficiently summary statistics over an entire exposed population as in the model of Wein et al. [2], for example, it is more convenient to use a single lognormal distribution, because numerical solutions to the convolution of two lognormal distributions, or a lognormal distribution with some other distribution for visit delay, is likely to make the overall computation inefficient or impractical.

We used data about visit delay for inhalational anthrax from a review of cases of inhalational anthrax by Holty [9]. We estimated the mu and sigma parameters of the lognormal distribution this data and maximum-likelihood estimation, with resulting values of 1.015 and 0.737, respectively.

Wu et al. [10] developed a method to approximate the distribution of the sum of two lognormal variables with a single lognormal distribution in the context of cellular phone technology. We apply this method to the incubation period and visit delay of inhalational anthrax, and provide a solution that anyone can use to quickly derive a single lognormal distribution for the incubation period plus visit delay. Our method also accounts for shorter incubation periods with higher doses of spores (per Wilkenings description of how various anthrax models incorporate this relationship [5]).

The method of Wu et al. begins by assuming that the two lognormal random variables are independent, and thus the moment-generating function of their sum is simply the product of two lognormal moment-

generating functions. Their method then approximates the moment-generating functions with their respective Gauss-Hermite representations. These approximations set up a system of equations whose solutions yield the mu and sigma parameters of the single lognormal distribution that approximates the distribution of the sum of the two lognormal variables.

In this work, we used the A1 model of inhalational anthrax of Wilkening [5] and our lognormal distribution fit to Holtys data. We assume that the dose of inhaled spores does not influence visit delay, and thus we assume a single distribution for visit delay over all spore doses. Note that the method of Wu et al. is general and we could apply it to other models of anthrax and models of other diseases.

We computed the mu and sigma parameters of the lognormal for the incubation period for log10(dose inhaled spores) = 1, 2, 3, 4, 5, 6, and 7 using the Wilkening A1 model, and input each of these distributions with the single distribution for visit delay into the procedure of Wu et al.. The end result was seven lognormal distributions representing the sum of the incubation period and visit delay at seven doses of inhaled spores. We computed the Hellinger distances between these lognormal approximations and the true convolutions and found that the Hellinger distances ranged from 0.0103 to 0.1021. We then used these 7 mu and sigma points and natural cubic spline interpolation to fit a curve to these mu and sigma parameters as a function of log10(spore dose), allowing one to easily compute a single lognormal distribution for the total time from spore exposure to the time of presentation to the healthcare system, for any given dose of inhaled spores.

In conclusion, our method for approximating the interval from exposure until the time an individual seeks medical care as the sum of two lognormal distributions--one for the incubation period and one for visit delay--is convenient, computationally efficient, and generates excellent approximations. Researchers in the future can build anthrax models that account for visit delay as well as incubation period without incurring the additional computational cost of modeling visit delay as a separate lognormal distribution. Moreover, the method is general applicable to lognormal incubation periods and visit delays of other diseases of interest.

References

1. Buckeridge DL, Burkom H, Moore A, Pavlin J, Cutchis P, Hogan W. Evaluation of syndromic surveillance systems--design of an epidemic simulation model. MMWR Morb Mortal Wkly Rep 2004;53 Suppl:137-43.

2. Wein LM, Craft DL, Kaplan EH. Emergency response to an anthrax attack. Proc Natl Acad Sci U S A 2003;100(7):4346-51.

3. Brookmeyer R, Johnson E, Bollinger R. Public health vaccination policies for containing an anthrax outbreak. Nature 2004;432(7019):901-4.

4. Webb GF, Blaser MJ. Mailborne transmission of anthrax: Modeling and implications. Proc Natl Acad Sci U S A 2002;99(10):7027-32.

5. Wilkening DA. Sverdlovsk revisited: modeling human inhalation anthrax. Proc Natl Acad Sci U S A 2006;103(20):7589-94.

6. Brookmeyer R, Johnson E, Barry S. Modeling the incubation period of anthrax. Stat Med 2005;24(4):531-42.

7. Sartwell PE. The distribution of incubation periods of infectious disease. 1949. Am J Epidemiol 1995;141(5):386-94.

8. Armenian H. Invited commentary on ,The distribution of incubation periods of infectious disease,. Am J Epidemiol 1995;141(5):385.

9. Holty JE, Bravata DM, Liu H, Olshen RA, McDonald KM, Owens DK. Systematic review: a century of inhalational anthrax cases from 1900 to 2005. Ann Intern Med 2006;144(4):270-80.

10. Wu J, Mehta NB, Zhang J. A flexible lognormal sum approximation method. In: IEEE Global Telecommunications Conference (GLOBECOM); 2005; 2005. p. 3413-3417.

A Simple, Versatile, Data-adaptive Approach for Alerting Based on Temporal Biosurveillance Data

Howard S. Burkom, (The Johns Hopkins University Applied Physics Laboratory), Howard.Burkom@jhuapl.edu, and
Sean Patrick Murphy, (The Johns Hopkins University Applied Physics Laboratory), Sean.Murphy@jhuapl.edu

Abstract: This effort describes a simple yet versatile method for automated data classification that addresses the problem of selecting appropriate alerting algorithms for biosurveillance data based on limited data history. This method is applicable to the univariate time series that result from syndromic classification of clinical records and also from nonclinical data such as filtered counts of over-the-counter remedy sales. Intended beneficiaries are local public health monitors using their own data streams as well as large system developers managing many disparate data types.

Numerous, recent papers have presented and evaluated algorithms for biosurveillance-related anomaly detection. However, authors of these papers are rarely able to share their datasets and can often publish only limited information describing them, so these papers do little to help a health monitor decide whether a published method will work well on the data at hand. Accentuating this problem, health monitors at 2005-2006 conferences and workshops related to automated health surveillance repeatedly expressed the need for modifiable case definitions and syndromic filters, thus obtaining time series whose behavior cannot be modeled in advance. Impromptu case definition changes may lead to changes in the scale and cyclic or seasonal series behavior. We demonstrate that mismatched algorithms and data can result in significant, systematic loss of sensitivity at practical false alarm rates. Therefore, the automated selection of suitable alerting methods is necessary.

Published alerting methods intended to control for trends and other systematic data behavior have used various regression-based models and other approaches such as wavelets and LMS filters [1-3]. The success of these methods is related to the presence of the day-of-week effects, annual trends, or other features that they are designed to model. A common approach to removing such expected features is the Phase I/Phase II paradigm of the statistical process control community. Control chart parameters, model features such as regression or filter coefficients, and sometimes alerting thresholds are calculated from a set of historic baseline data assumed to be representative of the data to be monitored. The inferred quantities are then applied for prospective surveillance. Because many time series monitored for biosurveillance are nonstationary, baseline-inferred parameters and thresholds may produce unexpected and uneven detection performance. Autoregressive methods and adaptive regression models and filters have been applied, as in [4, 5], to address this obstacle. Our approach utilizes prediction by generalized exponential smoothing, which we implement as a form of Holt-Winters (H-W) forecasting [6]. A comparison of this approach to nonadaptive and adaptive regression models yielded favorable results in [7] on multiple time series of two common types. The current effort discusses the automated selection of smoothing coefficients for H-W forecasting and the application of the H-W residuals in control charts. Smoothing coefficients yielding reliable daily forecasts may be obtained from a fairly small (as little as 2 months) representative data sample. We compute simple discriminants based on the scale, variability, overall trending, and day-of-week effects in the sample data to select from limited combinations of smoothing coefficients. The selection process may be wholly or partially overridden by user specifications based on knowledge of similar data.

The principal advantages of H-W forecasting are its capability to adapt to short-term trends without substantial model-fitting and its stability relative to the selected smoothing coefficients. Additionally, this approach can adapt to short-term trends without complex model-fitting, does not involve convergence problems, can be implemented in a spreadsheet, and can handle both rich and sparse data streams. Our particular H-W adaptations have been to account for day-of-week and holiday effects, avoid numerical

problems resulting from ongoing or temporary (due to data dropouts) sparseness, and avoid inappropriate training based on unexpected outliers. We evaluate the derived control charts and compare them to other alerting algorithms using receiver operating characteristic (ROC) curves based on realistic outbreak signals added to authentic data.

References

(1) Brillman JC, Burr T, Forslund D, Joyce E, Picard R and Umland E. Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance, BMC Medical Informatics and Decision Making 2005, 5:4, pp 1-14
http://www.biomedcentral.com/content/pdf/1472-6947-5-4.pdf

(2) Goldenberg A, Shmueli G, et al, Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales, Proc. Natl. Acad. Sci. USA, Vol. 99, Issue 8, 5237-5240, April 16, 2002

(3) Najmi AH, Magruder SF. An adaptive prediction and detection algorithm for multistream syndromic surveillance. BMC Med Inform Decis Mak. 2005 Oct 12; 5:33.

(4) Reis BY, Mandl KD, Time series modeling for syndromic surveillance (2003). BMC Medical Informatics and Decision Making 2003, 3:2

(5) Burkom, H.S., Development, Adaptation, and Assessment of Alerting Algorithms for Biosurveillance, Johns Hopkins APL Technical Digest 24, 4: 335-342.

(6) Chatfield C. The Holt-Winters Forecasting Procedure. Applied Statistics 1978; 27: 264-279

(7) Burkom, H., Murphy, S.P., and Shmueli G, Automated Time Series Forecasting for Biosurveillance, accepted for 2007 publication in Statistics in Medicine.

A Power Analysis of Two Surveillance Methods in Terms of Average Run Lengths

Gerald Shoultz, (Department of Statistics, Grand Valley State University), shoultzg@gvsu.edu,
Paul Stephenson, (Department of Statistics, Grand Valley State University), stephenp@gvsu.edu,
and
J. Wanzer Drane, (Department of Epidemiology and Biostatistics, University of South Carolina), wdrane@gwm.sc.edu

Abstract: This talk compares two methods for testing hypothesis usable in disease surveillance and process control to determine which is more powerful: TEXAS (Hardy et al 1980) and CUSUM (Hawkins and Olwell 1998). TEXAS, a modification of the procedures of Shewhart (1931), uses a two-step decision rule to determine when a process is out-of-control. CUSUM finds a process to be out-of-control when the sum of a set of measurements exceeds a given threshold. While there are many reasons for monitoring disease incidence in a community or region, one likely application of such methodology is identifying if a government agency should investigate whether or not a terrorist event has occurred. First, the authors will discuss how these process control procedures can be used to monitor disease surveillance. Then the authors will present a simulation that compares the performance of the TEXAS and CUSUM methods to determine which method is more powerful for a variety of hypotheses.

Thursday, 1:30 – 3:00 Track D - **Contributed Session #5:** *"Data Analysis - II"*
Chair: David Marchette

Factors that Influence Algorithm Performance in the Face Recognition Grand Challenge
Jonathon Phillips, (NIST), jonathon@nist.gov
Ross Beveridge (Colorado State University),
Geof H. Givens (Colorado State University),
and Bruce A. Draper (Colorado State University)

Abstract: Over the last two decades the effort to develop effective automatic face recognition has resulted in hundreds, if not thousands of papers. Typically, these papers report performance on a single data set in order to draw comparisons between alternative approaches. This type of analysis is valuable when the goal is to conclude that a particular approach is superior to another on a very specific task as exemplified by the data set. However, this style of analysis tells us little about underlying factors that make recognition easier or harder. When it is addressed at all, the question of what factors affect recognition performance is almost invariably addressed by dataset partitioning. Consider studying the effects of pose and illumination. Several carefully constructed datasets have been developed that lend themselves to studying these factors with partitioned data. Work with the Yale data set and subsequently the PIE data set typically falls into this category. In other words, studies look at relative performance across changes in pose, illumination, or both as exemplified by performance on distinct data partitions. Partitioned data set analysis has also been applied to the question of whether women or men are easier to recognize.

Unfortunately, this approach is less effective the moment one begins to ask questions about more than a handful of factors. It is also far less practical, due to the combinatorial explosion of partitions over multiple factors. Of greater importance is the fact that the practical limitations of partitioning impose important limitations on the ability to control for confounding factors. If one skirts around the combinatorial problem by resorting to marginal analysis (i.e., abandoning control via partitioning), control of confounding effects is eliminated altogether. More sophisticated multi-factor statistical techniques provide greater control and permit more thorough evaluation of factor effects.

Generalized linear mixed models (GLMMs) are one such technique. This paper uses GLMMs to provide the largest statistical analysis to date of factors that influence face recognition performance. Our analysis investigates how a set of 12 factors, henceforth called covariates, predict verification rate at various false accept rates. Performance data for three algorithms from the Face Recognition Grand Challenge Experiment 4 are used in our analysis. The covariates include such things as gender, race, age, distance between eyes in pixels, and apparent focus of the imagery.

Our analysis shows that covariates have a significant effect on performance and that overall performance reported in evaluations does not give a complete picture of the performance properties of face recognition algorithms. The analysis shows for the first time the effect of image based covariates such as size, tilt, and focus of a face on performance.

Our analysis shows that a number of the assumptions of the automatic face recognition community are wrong. Our analysis also provides additional scientific evidence supporting observed effects of subject covariates on performance.

Splatter Terrain: An Interactive 3D Visualization Framework for Understanding Dense Scatterplots
Pranab K. Banerjee, (Space Dynamics Laboratory), Pranab.Banerjee@sdl.usu.edu

Abstract: The information age has brought with it the challenge of managing data deluge. Today, we are surrounded by more data than we can comprehend and this gap between the technological capabilities in data acquisition and information extraction seems to be widening. The field of sensor hardware development has seen tremendous growth in recent years fueled by our desire to understand and investigate the world around us in greater detail, in multiple modalities, and from diverse viewpoints. These

developments have been beneficial to defense and national security since newer sensors can capture data at higher resolutions, sense more channels, and handle higher communication bandwidth for data sharing, thus providing the data streams necessary for real-time high fidelity intelligence gathering and situational awareness. However, data does not equate to information. Domain specific relevant information embedded in high volume, high velocity raw data streams is often sparse and it requires careful data mining and analysis to discover and comprehend such knowledge. The enormous volume, velocity, variety, and high dimensionality of data produced by modern data acquisition machinery overwhelm our capabilities to analyze and comprehend such embedded information because of computational limitations, as well as human cognitive constraints.

Effective comprehension of trends, outliers, and correlations in data streams is important for gaining crucial situational awareness for timely decision making. This is particularly important in defense and national security where the utility and effectiveness of extracted intelligence may have a short utility span and real-time or near real-time comprehension of embedded information is critical for optimal exploitation of this intelligence. Visualization can play a key role in such data analysis and knowledge discovery since the faculty of human visual cognition has evolved to be an efficient, massively parallel and fairly robust pattern discovery and recognition engine capable of identifying interesting visual features even in the presence of some noise. But, many time-proven visualization tools that are effective for smaller data sets break down when applied to large data volumes. As a result, innovative techniques and algorithms are needed for large dataset visualization.

A particularly useful and time proven method for discovery of correlation and outliers in data is the scatterplot which is a two dimensional x-y point plot where each axis represents an entity of interest and a point in the plot corresponds to an (x,y) tuple that appears in a record in the dataset. Visual inspection of these plots can reveal the relationships between these entities. Traditional 2D scatterplots, however, suffer from a serious drawback for large datasets - that of visual clutter resulting from spatial overlapping of points, making them unresolvable. The clutter mitigation techniques proposed in the past can be broadly classified into two categories: (i) data reduction, and (ii) spatial reorganization.

Techniques in the data reduction category are essentially based on various sampling or quantization schemes that result in fewer points to be displayed. Uniform random sampling can produce a low density representation that can maintain the overall trend in the data. However, sampling can introduce artifacts not present in original data and relatively small clusters may not be preserved. Besides, the process of crossing resolution boundaries in multi-resolution representation spaces resulting from various levels of sampling in order to explore the details as well as overviews can pose perceptual continuity challenges.

Non-uniform sampling techniques can address the issue of preserving small clusters in the lower density representation, but these schemes are undesirable for scatterplots because they alter the underlying statistical properties of the dataset.

Data filtering schemes have been proposed to reduce clutter by selectively displaying certain subsets of the original data space according to some selection or filtering criteria. For example, visual clustering schemes reduce clutter by aggregating pixels that are similar, based on a predefined similarity metric. However, these are not particularly suitable in the case of scatterplots because they alter the underlying data, and any correlation information is lost in the process. Techniques based on distortion of visual representation space, such as the ``fisheye view'' are useful in clutter reduction in many information visualization tasks. It displays a low density representation of the overall data space but allows higher resolution views of local regions of interest. However, this approach is not helpful for scatterplots because visual discovery of correlations and clusters between the two dimensions become difficult and cognitively stressful unless the entire display has a uniform resolution.

Spatial reorganization techniques are based on either a heuristic or an optimal bijective spatial mapping function that redefines the spatial coordinates of the points. The main drawback of this class of algorithms is that the number of points that can be effectively displayed is limited by the number of pixels in the display.

This paper describes a novel fully interactive and intuitive 3D environment for effective visualization and analysis of dense scatterplots. Conventional 2D scatterplots are transformed to a 3D terrain, called Splatter Terrain, by adapting the idea of splatting from the field of 3D volume rendering. This technique does not require any data reduction or data perturbation and it produces a visually intuitive and clutter free overview of the point densities for easy identification of interesting regions for further drill down analysis. This approach is motivated by Ben Schneiderman's visual information seeking mantra: ``Overview first, zoom and filter, then details-on-demand".

To generate the Splatter Terrain, each point in the scatterplot is subjected to a splatting kernel that has the effect of distributing the influence of the point to its immediate neighborhood. A 2D image, called the ``splat image" is generated where each pixel corresponds to the sum of the influences from all neighboring points affecting that spatial location. The splatting process, thus, converts a scatterplot consisting of discrete distribution of points to a 2D image covering the spatial extent of the plot. This image is then used as a height-map to render the final 3D Splatter Terrain that makes it easy to visually discover underlying statistical relationships between the parameters. The height of the terrain at a point corresponds to the number of points in the neighborhood around that spatial location. A 2D Gaussian is commonly used as the splatting kernel and the influence of a point effected by the kernel is limited to a finite domain centered around the point for computational efficiency. For faster approximate computation of the 2D splat image, a regular grid of appropriate resolution spanning the entire scatter plot is generated where original points in the scatter plot are binned into appropriate grid locations. The splat image is then obtained as the convolution of the splatting kernel with the grid bins carried out in the frequency domain. The Splatter Terrain is texture mapped with a pseudo-colored version of the splat image for enhanced comprehension of the point distribution in the scatter plot.

The Splatter Terrain provides an easy to understand overview of an arbitrarily dense and large scatterplot but it is not good at showing outliers or small clusters as they may be hard to distinguish from a flat region. To address this issue, the visualization tool displays the original 2D scatter plot along with the Splatter Terrain in a spatially registered manner. A pair of interactive orthonormal semi-transparent planes make it easy to visually associate any point on the surface of the Splatter Terrain to a location in the scatterplot for on-demand drill down analysis. A heads-up display scheme is employed to show the coordinates of the parameter space as well as point density information dynamically as these planes are interactively manipulated to select points on the Splatter Terrain surface. This allows the user to view statistical information without moving his/her eyes away from the core visual representation. A common problem in 3D visualization is occlusion. The tool addresses this problem by providing a fully interactive environment that allows the user to manipulate the terrain and the spatially registered scatterplot in real time through rotation and zooming. In addition, a pair of semi-transparent parallel ``measuring planes" are deployed for easy comparison of the heights of the Splatter Terrain at two different locations. This is useful for exploring regions with subtle differences in point densities that may be hard to comprehend from the pseudo-colormapped splat image or iso-contours. The measuring planes can be texture mapped with the splat image with a user defined colormap and their levels of transparency can be interactively varied.

The visualization framework has been tested with atmospheric datasets showing significant clutter in the scatterplot, and the initial response has been positive based on an informal user study.

Ship Itinerary Predictability

Patricia H. Carter, (Naval Surface Warfare Center), `patricia.h.carter@navy.mil`

Abstract: Understanding the nature of ship itineraries is a first step towards modeling commercial ship behavior to support Maritime Domain Awareness. The goal is to determine patterns of individual and collective commercial vessel behavior, to establish normal behavior patterns and to accurately detect anomalous behavior. Methods to produce measures of anomaly will facilitate risk evaluation of vessels of interest in tactical situations.

A core model of ship behavior that is susceptible to mathematical and statistical analysis is the ship itinerary. This model can accommodate the association of additional data about the ship, ports and the local

and global situation. A further abstraction of the ship itinerary is the sequence of ports visited. Sequences of port visits is the model investigated in this paper. The scope of global shipping traffic is vast: more than 100,000 ships and several thousands of ports. There is a huge amount of variability in the predictability of ship itinerary patterns. The first question to ask is how predictable and regular are individual ship itineraries? In other words, what are the normal patterns? The kinds of patterns one might expect include itineraries that are 1) periodic sequences, 2) sequences that contain periodic subsequences with insertions and deletions, 3) sequences with multiple periodic subsequences, 4) sequences with disordered and non-repeating behavior. Sequence processing is a well developed discipline as it is fundamental in computer science, text processing and bioinformatics. There are many different measures of periodicity and recurrence that can be used to quantify how close a sequence is to exhibiting regular, repeating behavior. These can be used to explore whether the behavior can be clustered into types or if there is a continuous spectrum of behavior. The predictability of port sequences on a ship's itinerary can be investigated by assigning a measure of anomaly to port visits. For each ship, given a sequence of past port visits, one predicts how likely each of a number of choices of ports is the next port to be visited; here a weighted all-gram profile approach is used with the principle of maximizing the recurrence properties of the sequence. An anomaly score is then determined by comparing the actual port visited with its prediction weight. The distribution of anomaly scores that occur over a set of ship itineraries is indicative of the predictability of port sequences.

This approach will be evaluated by applying it to a data set consisting of a little over a year's schedules from about 2000 commercial ships visiting about 700 ports, for a total of 137,000 port visits.

Future and ongoing work addresses the more difficult problem of modeling global behavior. The collective patterns of behavior are a product of a system which is self-organizing rather than hierarchically determined and emerging patterns are both driven and constrained by a myriad of forces and factors. An appropriate starting point would be a dynamic graph model. A limiting factor in the analysis of collective patterns and the development of appropriate models is the difficulty and expense of assembling a sufficient amount of data, because most of the data is owned by private entities. Understanding collective commercial vessel behavior and its trends is critical in the construction of a global maritime situational picture.

The dynamic behavior that is captured by port sequence analysis lies between the micro-scale behavior involved in near real time ship tracking and the macro-scale behavior of the global commercial shipping system. In this paper the utility of various sequence analysis methods applied to port sequence data is explored; these methods are appropriate and useful for the meso-scale behavior but different tools will be required for the micro- and macro-scale behavior of commercial shipping traffic.

Thursday, 3:00 – 3:30 BREAK

Thursday, 3:30 – 5:00 Track C - **Contributed Session #6:** *"Survey and Modeling"*
Chair: Yasmin Said

A Post-Katrina Survey in New Orleans: Sampling the Unsettled
David Banks, (Duke University), banks@stat.duke.edu

Abstract: After Katrina, the NSF funded a number of projects to study the impact of the disaster and the issues affecting recovery. This talk describes one such project, a survey of Katrina refugees to identify the factors that influenced their decision on whether or not to evacuate before the storm struck, and the factors that affected the quality of their post-evacuation experience. As part of this project, the researchers learned a number of lessons that should usefully inform other survey efforts undertaken in the context of disaster response.

Methods for Ensuring that Statistical Information does not Reveal Underlying Individual Data
Paul B. Massell, (U.S. Census Bureau), paul.b.massell@census.gov

Abstract: Many federal agencies collect information from individual persons, households, and companies with the goal of producing data products which represent statistical summaries of the individuals records,

often in the form of tables but sometimes in some other form, such as statistical models. Another type of data product occasionally released is microdata, a subset of the records collected in which, typically, key identifying fields are eliminated and other fields are smoothed to prevent their use in an identification. Even when only statistical summaries are released, the agency must take care to protect the confidentiality of the underlying individual data. This is needed, for example, when tables have cells that represent totals (e.g., income) based on a very small number of individuals. We give a brief overview of methods used for protecting the individuals data underlying statistical tables. We conjecture that such methods may be useful in the development of statistical (profiling) models for identifying bad guys in which there may be a strong legal and/or national security interest in not revealing any of the underlying individuals data.

The Importance of Dedicated Experiments to Support Validation and Calibration Activities
Genetha Gray, (Sandia National Labs), gagray@sandia.gov,
Monica Martinez-Canales, (Sandia National Labs), mmarti7@sandia,gov,
Cheryl Lam, (Sandia National Labs), clam@sandia,gov, and
Brian Owens, (Sandia National Labs), beowens@sandia.gov

Abstract: In recent years, numerical modeling and simulation have been used to augment and replace physical experiments in the study and design of complex engineering and physics systems. Moreover, the results of these simulations are often considered by decision makers in areas such as defense and national security. Therefore, validation and verification (V&V) activities have become critical for determining simulation-based confidence and predictive capabilities. For example, code verification must be used to confirm that the underlying equations are being solved correctly. In addition, validation processes should be applied to answer questions of correctness of the equations for the phenomena being modeled and the application being studied. Moreover, validation metrics must be carefully chosen in order to explicitly compare experimental and computational results and quantify the uncertainties in these comparisons.

Data is the driver of the V&V process. While existing experimental data is often used for validation activities, in some cases, existing data may be inadequate or inappropriate for comprehensive validation. For example, data may only be available in limited quantities or data sets might not be replicated making quantification of measurement variability within experimental factors impossible. Selecting new experiments requires tools that elucidate the behavior being studied as well as how it will be tested. The goal is to cover the space created by the experimental factors of importance as well as to test at extremes of testable space to ensure confidence in extrapolating the model to untestable regions. The variety of experimental conditions, experimental measurement errors, and part-to-part variation must also be considered.

Overall, the V&V process for modeling and simulation can provide the best estimates of what can happen and the likelihood of it happening when uncertainties are taken into account. In order to carry out the validation activities, experiments must be carefully planned and executed to provide adequate and appropriate data. We will describe this process for the validation activities related to Xyce, an electrical circuit simulator developed at Sandia. We will also discuss how V&V can be applied to the processes of decision making and risk analysis.

Thursday, 3:30 – 5:00 Track D - **Contributed Session #7:** *"Anomaly Detection"*
Chair: Barton Clark

Anomaly Detection in Space-Time (and higher dimensional) Point Processes
Michael D. Porter, (NCSU), porter@stat.ncsu.edu

Abstract: There is a growing need to develop methodologies for change detection in space-time processes. This talk discusses some approaches to anomaly detection (a specific type of change where the change occurs in a local region of space) in space-time point processes. The problem of detecting such changes is applicable in areas such as disease surveillance, computer intrusion detection, target detection, and crime and terrorism.

We take a likelihood based approach where the unknown pre and post change parameters are estimated adaptively, thus expanding the common GLR, CUMSUM, and Shiryaev-Roberts change detection methodologies. As one of the post-change parameters is the region where change has occurred, we also discuss some methods to identify this region in 2-D and higher dimensional spaces.

Bayesian spatial scan statistic adjusted for overdispersion and spatial correlation.

Deepak Agrawal, (Yahoo! Research), dagarwal@gmail.com

Abstract: Spatial scan statistic has become the method of choice for detecting spatial clustering after adjusting for inhomogeneity. The method is particularly suitable in applications where the goal is to find the actual location of spatial clusters or ``hotspots'' as opposed to testing for global clustering. The method has been extremely successful and has found applications in diverse areas ranging from biosurveillance, forestry, criminology, psychology etc. The method proceeds by scanning the study region using all possible spatial sub-regions that conform to some geometric shape (e.g., circle, rectangle, ellipsoid, etc). Each sub-region is assigned a discrepancy measure which is based on a likelihood ratio test that compares the intensity inside the sub-region with the intensity outside. The sub-region with the maximum discrepancy is generally declared to be a ``hotspot'' provided it is statistically significant. The significance test is based on an expensive randomization procedure which computes a Monte Carlo p-value by repeatedly (approximately 10K times) generating realizations under the null hypothesis of no spatial clustering.

In this paper, we propose a Bayesian solution to the problem. A Bayesian solution has several advantages in this scenario. First, hotspot detection is based on posterior probabilities of models corresponding to each sub-region and hence there is no need to conduct the randomization procedure. This gain in computational efficiency is obtained by performing a slightly more expensive discrepancy calculation for each sub-region wherein a simple and closed form likelihood maximization is substituted by a numerical integration routine. Second, compared to the classical approach where multiple hotspots are generally detected using a conservative test, detecting multiple hotspots in the Bayesian framework is automatic and does not require any additional machinery. Finally, the Bayesian setting also provides a natural framework to incorporate any prior knowledge that might be known about the hotspots. To the best of our knowledge, no rigorous work in a Bayesian framework exists in the statistics literature. Recently, a Bayesian solution to the problem was proposed by (Neil et al., NIPS 2005) in the machine learning literature. However, their solution made strong assumptions on the priors of sub-regions. Moreover, it is not possible to adjust for additional characteristics like overdispersion and spatial correlation using their framework. Such adjustments are potentially useful in the context of biosurveillance where the analyst might not be interested in investigating clusters that are caused only due to presence of routine overdispersion relative to the usual Poisson or Bernoulli model. Adjusting for such routine characteristics in the baseline model can potentially reduce false positives and enhance disease monitoring systems used in public health.

Our contributions are in two directions. First, we propose a modeling framework using a point process formulation. We propose the use of a Cox process to enhance the usual assumptions of a Poisson process. The Cox process assumes that conditional on a latent error process, data comes from a Poisson process. Marginalizing over the latent process enables adjusting for features like overdispersion that might be present in the data. For instance, assuming a gamma distribution gives rise to the usual negative binomial distribution that has been widely used to model overdispersion in count data. Other possibilities include a Conditionally Autoregressive Process (CAR) that is widely used to model spatial correlation in epidemiology. Next, we provide a Bayesian solution to the problem in our proposed framework. Our solution does not depend on eliciting data based priors for each sub-region as in (Neil et al., NIPS 2005). In fact, the main computational bottleneck in the Bayesian procedure is the computation of a Bayes factor for each sub-region. For the usual Poisson model proposed by (Kulldorf, 1997), this boils down to computing a 2-dimensional integral for each sub-region which is done efficiently and accurately using a Laplace approximation. For a negative binomial model, the same strategy works with the 2-d integral being replaced by a 3-d integral. For models like CAR that are multivariate in nature, one needs to compute a high dimensional integral. The Laplace approximation does not provide accurate answers in this scenario and one needs to take recourse to computationally intensive procedures like MCMC. However, the computations are amenable to parallel computing and could be performed efficiently in a cluster computing

environment for reasonably sized datasets. We illustrate the efficacy of our procedure on datasets that have been previously analyzed in the literature.

Using Scan Statistics for Anomaly Detection in Genetic Regulatory Networks

Christopher C. Overall, (George Mason University), coverall@gmu.edu,
Jeffrey L. Solka, (Dahlgren Division of the Naval Surface Warfare Center), Jeffrey.Solka@navy.mil,
Jennifer W. Weller, (George Mason University), jweller@gmu.edu, and
Carey E. Priebe, (Johns Hopkins University), cep@jhu.edu

Abstract: Biological systems contain many levels of complex interactions between heterogeneous components, the dynamics of which are usually non-linear. Each functional layer forms an interacting network, and the network layers interact, but the components are not completely connected. It has become increasingly popular to represent these biological interactions as a network (graph) in which a node (vertex) represents a biological molecule or functional complex and an edge represents a relationship between the two molecules. This graph representation provides a powerful and intuitive framework because the full power of graph theory can be harnessed for analyzing the global behavior of the system, but it comes with a price; the dynamics of the system are lost when the interactions are represented as a static graph. In general, biological researchers often want to determine if and when the relationship between biological entities alters significantly over time and in response to the environment in the system under study. In other words, biologists are interested when and where an anomaly occurs and this requires that the dynamics of the system be incorporated into the analysis.

The problem of detecting anomalies in biological networks is analogous to anomaly-based network intrusion detection in the computer network security domain. There is a large amount of data that requires automated techniques for determining normal network behavior and then using this prior history to determine when an anomalous change has occurred at one or more nodes in the network. These techniques detect anomalous behavior in the network that might not have been deduced by a human, allowing the analyst or researcher to hone in on the anomaly and to determine if it is significant. Although many anomaly-based network intrusion detection techniques have been developed for computer networks, to our knowledge, there are not any similar anomaly detection techniques for biological networks.

We have developed a technique for anomaly detection in genetic networks that are generated from time-series transcriptional profiling experiments, of the type that are measured on microarray platforms or RT-PCR devices, and successfully applied it to a time-series Drosophila microarray dataset. The technique is a hybrid solution for the study of biological interaction networks, incorporating some of the dynamics of the system while using the simplifying network representation as the analysis framework. First, a genetic regulatory network is constructed for the genes under investigation. Then univariate and/or multivariate model-based clustering is used to create a time sequence of graphs using the time-series gene expression dataset and the genetic regulatory network. Finally, the series of graphs is analyzed for anomalies in gene activity over time using the graph-based scan statistics of Priebe et al. (2006).

Although the characterization of static biological interaction networks and the interplay between them is far from complete, even for model organisms, we feel that it is nevertheless important to take the next logical step and begin to explore techniques for automated anomaly detection in these networks. We hope that our anomaly detection methodology will spur interest in, and appreciation for, this type of analysis in the biological sciences.

Author Index