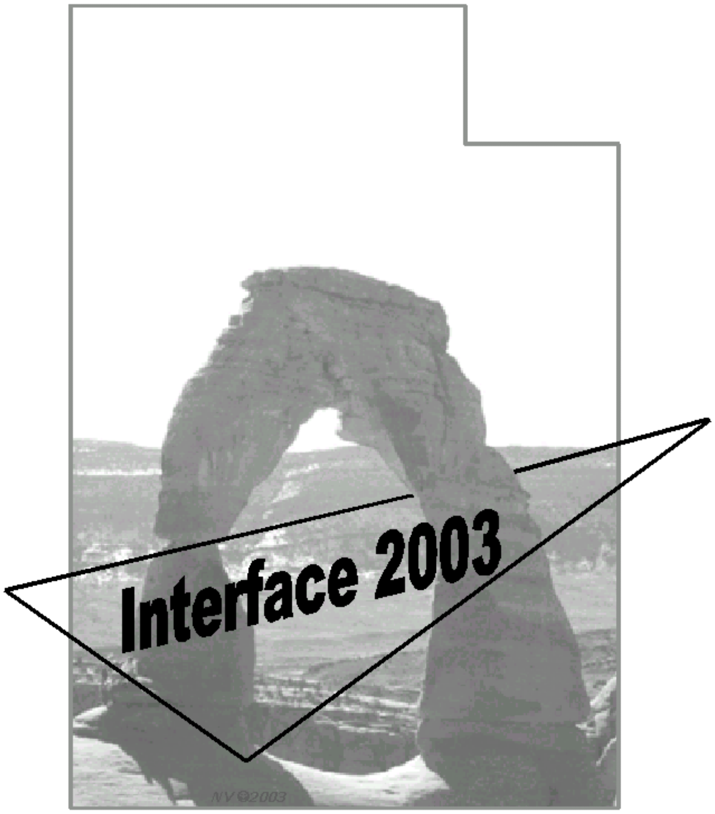
# 35TH SYMPOSIUM ON THE INTERFACE: COMPUTING SCIENCE AND STATISTICS

## SECURITY AND INFRASTRUCTURE PROTECTION

March 12–15, 2003
Sheraton City Centre
Salt Lake City, Utah



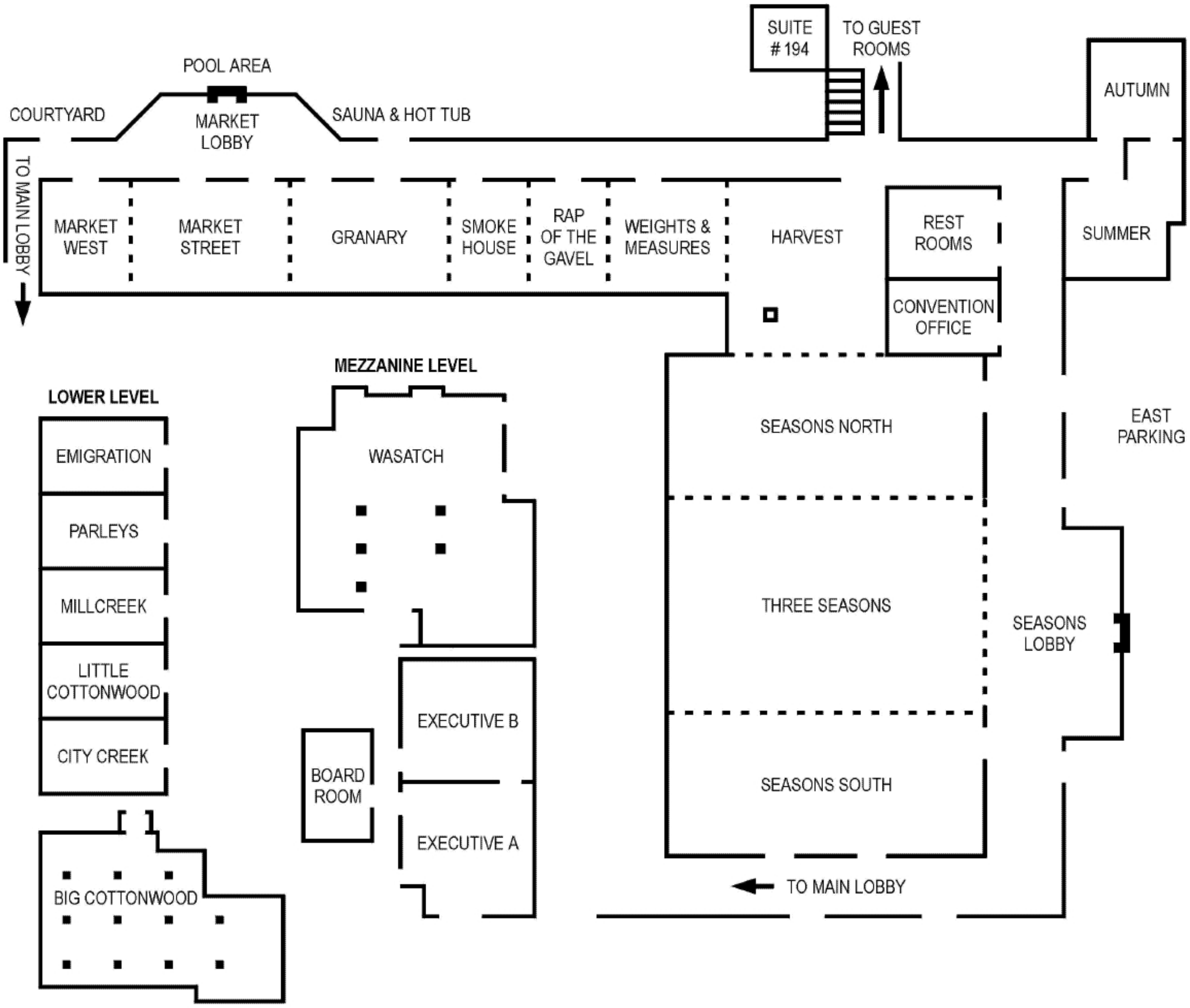
Interface Foundation Of North America

Hosted by:
Utah State University
Co-Chairs: Michael C. Minnotte and Jürgen Symanzik

Sponsored by:
Army Research Office
Office of Naval Research
National Security Agency
ASA Section on Statistical Computing
ASA Section on Statistical Graphics

Cooperating Organizations:
ASA, CSNA, ENAR, IASC, IMS
INFORMS, SIAM, WNAR

## Sheraton City Centre Conference Facilities



# Interface 2003

## Sponsor

The Interface Foundation of North America (IFNA) is a nonprofit education corporation founded in 1987 to sponsor the symposium and publish the proceedings. IFNA also co-publishes the *Journal of Computational and Graphical Statistics.*

## Business Office

Interface Foundation of North America
P.O. Box 7460
Fairfax, VA 22039-7460
(703) 993-4635
`interface@galaxy.gmu.edu`

## Symposium Co-Chairs

Michael C. Minnotte
Department of Mathematics and Statistics
Utah State University
(435) 797-2844
`minnotte@math.usu.edu`

Jürgen Symanzik
Department of Mathematics and Statistics
Utah State University
(435) 797-0696
`symanzik@math.usu.edu`

## Proceedings Editors

Michael Minnotte
Jürgen Symanzik
Ed Wegman

Proceedings papers and all supporting materials should be sent by July 1, 2003, electronically to `iface03@math.usu.edu`. Please see page 19 for full instructions.

## Conference Arrangements

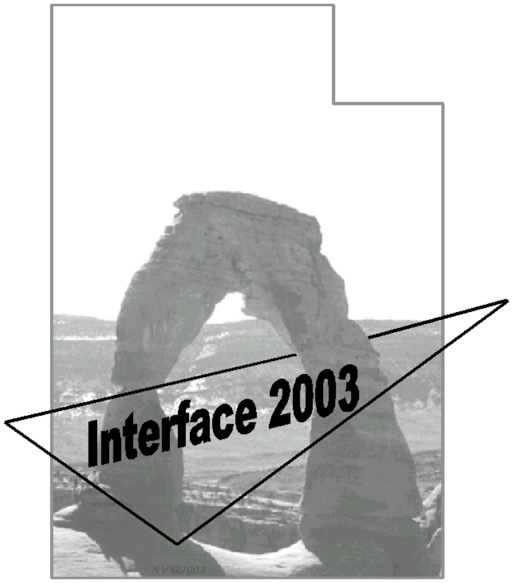
Michael Minnotte, Jürgen Symanzik, Ed Wegman, Liz Quigley, Natascha Vukasinovic, Adele Cutler, Jeff Solka, David Marchette, Andrejus Parfionovas, Kady Schneiter, Qian Zhao, Rena Gilden, Nancy Smart, the Utah State University Department of Mathematics and Statistics

## Program Committee

Michael Minnotte (Utah State University)
Jürgen Symanzik (Utah State University)
Dale Anderson (PNNL)
Sue Bell (National Cancer Institute)
Barry Bodt (U.S. Army Research Laboratory)
Amy Braverman (Jet Propulsion Laboratory)
Daniel Carr (George Mason University)
Steve Cohen (Bureau of Labor Statistics)
Adele Cutler (Utah State University)
Catherine Dippo (Bureau of Labor Statistics)
Robert Edsall (Arizona State University)
Robert Gentleman (Harvard University)
Jimmie Givens (Nat. Center for Health Stats.)
Arnold Goodman (University of California, Irvine)
Scott Grimshaw (Brigham Young University)
John Hinde (Nat. Univ. of Ireland, Galway)
Alan Karr (NISS)
Soumendra Lahiri (Iowa State University)
David Marchette (NSWC)
Steve Marron (University of North Carolina)
Gretchen Moisen (US Forest Service)
Silvia Nittel (University of Maine)
Linda Pickle (National Cancer Institute)
David Scott (Rice University)
Jeff Solka (Naval Surface Warfare Center)
Shailaja Suryawanshi (Merck)
Deborah Swayne (AT&T)
Bill Szewczyk (National Security Agency)
Duncan Temple Lang (Bell Labs, Lucent)
Sandra Thompson (PNNL)
Natascha Vukasinovic (Monsanto Company)
Ed Wegman (George Mason University)
Andrew Westlake (Survey & Stat. Computing)
Sharon Wunschel (PNNL)

## Contacts

Interface 2003
Department of Mathematics and Statistics
Utah State University
Logan UT 84322-3900
`iface03@math.usu.edu`
`http://www.math.usu.edu/∼iface03`
Fax: (435) 797-1822



## Table of Contents

# Interface 2003 Program

## Wednesday, March 12, 2003

| | |
|---|---|
| 7:30am–7:00pm | *Harvest Room*<br>Registration and Exhibits |
| 8:00am–12:00pm | *Weights & Measures Room*<br>Short Course: Statistical Methods in Computer Security |
| 1:30pm–5:30pm | *Weights & Measures Room*<br>Short Course: Topics in Computational Statistics with MATLAB |
| 6:00pm–8:00pm | *Summer Room*<br>Interface Board of Governors Meeting (closed) |
| 8:00pm–10:00pm | *Seasons North Room*<br>Reception and Mixer |

## Thursday, March 13, 2003

| | | | | | |
|---|---|---|---|---|---|
| 7:00am–5:00pm | *Harvest Room*<br>Registration and Exhibits | | | | |
| 7:30am–8:30am | *Harvest Room*<br>Continental Breakfast | | | | |
| 8:30am–9:45am | *Three Seasons Room*<br>Opening Remarks and Keynote Address | | | | |
| 9:45am–10:30am | *Harvest Room*<br>Coffee Break | | | | |
| 10:30am–12:15pm | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* | *Market West Room* |
| | R (invited) | Bioinformatics (invited) | Security and Infrastructure Protection (invited) | Visualization Tools for Health Statistics (invited) | Statistical Graphs (contributed) |
| 12:15pm–1:45pm | Lunch Break | | | | |
| 1:45pm–3:30pm | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* | *Market West Room* |
| | Financial Risk and Fraud Detection (invited) | Statistical Methods in Genetics (invited) | Analysis and Modelling of Internet Traffic (invited) | Analysis and Visualization of Graph Data (invited) | Computational Methods (contributed) |
| 3:30pm–4:00pm | *Harvest Room*<br>Refreshment Break | | | | |
| 4:00pm–5:45pm | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* | |
| | Face and Pattern Recognition (invited) | Computer-Intensive Methods (invited) | Computational Statistics and NSA (invited) | Graphics and Visualization (contributed) | |
| 7:00pm–10:00pm | *Three Seasons Room*<br>Banquet | | | | |

## Friday, March 14, 2003

| | | | | | |
|---|---|---|---|---|---|
| 7:30am–5:00pm | *Harvest Room*<br>Registration and Exhibits | | | | |
| 7:30am–8:00am | *Harvest Room*<br>Continental Breakfast | | | | |
| 8:00am–9:45am | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* | *Market West Room* |
| | Sensors for Biological Threats (invited) | Best of KDD-2002 (invited) | Homeland Security (invited) | Modern Text Processing (invited) | Environmental Statistics (contributed) |
| 9:45am–10:30am | *Harvest Room*<br>Coffee Break | | | | |
| 10:30am–12:15pm | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* | *Market West Room* |
| | Social Networks (invited) | Best of IASC (invited) | Digital Government (invited) | Smoothing and Feature Detection (invited) | Safety and Security (contributed) |
| 12:15pm–1:45pm | Lunch Break | | | | |
| 1:45pm–3:30pm | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* | *Market West Room* |
| | Public Health Preparedness (invited) | Best of *JCGS* (invited) | Infrastructure Security (invited) | Statistical Computing (refereed) | Bioinformatics (contributed) |
| 3:30pm–4:00pm | *Harvest Room*<br>Refreshment Break | | | | |
| 4:00pm–5:45pm | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* | |
| | Computer Security (invited) | Interactive GeoGraphics for the Web (invited) | Design and Computation (contributed) | Nonparametrics (refereed) | |

## Saturday, March 15, 2003

| | | | | |
|---|---|---|---|---|
| 7:30am–11:00am | *Harvest Room*<br>Registration and Exhibits | | | |
| 7:30am–8:00am | *Harvest Room*<br>Continental Breakfast | | | |
| 8:00am–9:45am | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* |
| | Statistical Data Bases (invited) | Surveillance (invited) | Prediction of Catastrophic Events (invited) | Graphics for Bio and Chem Informatics (invited) |
| 9:45am–10:30am | *Harvest Room*<br>Coffee Break | | | |
| 10:30am–12:15pm | *Weights & Measures Room* | *Smokehouse/ Rap of the Gavel Room* | *Granary Room* | *Market Street Room* |
| | Large Databases and Data Streams (invited) | Data Mining Combat Simulations (invited) | Forests at Risk (invited) | Nonparametric Methods and Applications (contributed) |

# Interface 2003 Conference Schedule

## Wednesday, March 12, 2003

7:30 a.m. — **Registration begins** (Harvest Room)

8:00 a.m. to 5:30 p.m. — Interface Half-day Short Courses (Weights & Measures Room):

8:00 a.m. to 12:00 noon — **Short Course I: Statistical Methods in Computer Security**, David Marchette, Naval Surface Warfare Center

1:30 p.m. to 5:30 p.m. — **Short Course II: Topics in Computational Statistics with MATLAB**, Jeffrey L. Solka, Naval Surface Warfare Center

6:00 p.m. to 8:00 p.m. — **Interface Board of Governors Meeting** (Summer Room, closed)

8:00 p.m. to 10:00 p.m. — **Reception and Mixer** (Seasons North Room)

## Thursday, March 13, 2003

7:00 a.m. — **Registration Continues** (Harvest Room)

8:30 a.m. — **Opening Remarks** (Three Seasons Room)

8:45 a.m. — **Keynote Address** (Three Seasons Room): *Considerations of Inspection for Homeland Security with Cross Linkages to Quality Control, Game Theory, and Stochastic Simulation*, James R. Thompson, Rice University

10:30 a.m to 12:15 p.m. — **R** (Weights & Measures Room — Invited Session) Organizers: Robert Gentleman, Harvard University, and Duncan Temple Lang, Bell Labs, Lucent Technologies. Session Chair: Robert Gentleman, Harvard University

10:30 a.m. — *Technology Transfer in Industry using an R-COM Interface*, Scott Vander Wiel, John Chambers, Suresh Goyal, David James, and Duncan Temple Lang, Bell Labs, Lucent Technologies

11:00 a.m. — *Inter-System Interfaces for S*, Duncan Temple Lang, Bell Laboratories, Lucent Technologies

11:30 a.m. — *Sparse Linear Algebra for the R Language*, Pin Ng, Northern Arizona University, and Roger Koenker, University of Illinois

12:00 noon — Discussant: Robert Gentleman, Harvard University

---

10:30 a.m to 12:15 p.m. — **Bioinformatics** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: Adele Cutler, Utah State University

10:30 a.m. — *Exploratory Detection of Differential Gene Expression*, John D. Storey, University of California, Berkeley

11:05 a.m. — *Statistical Issues in the Design of Microarray Experiments*, Jean Yee Hwa Yang, UCSF , and Terry Speed, UC Berkeley

11:40 a.m. — *Exploration, Normalization, Summaries and Software of Affymetrix GeneChip Probe Level Data*, Rafael Irizarry, Johns Hopkins

---

10:30 a.m to 12:15 p.m. — **Security & Infrastructure Protection: Interface of Computing Science and Statistics to the Rescue** (Granary Room — Invited Session) Organizer and Session Chair: Arnold Goodman, University of California, Irvine

10:30 a.m. — *Statistical Opportunities in Network Security*, David Marchette, NSWC

11:05 a.m. — *Using Statistics to Detect and Thwart Denial of Service Attacks*, Carla Brodley, Purdue University

11:40 a.m. — *Maximizing Quality and Value in Security: Challenges for Computer Scientists, Statisticians and their Clients*, Arnold Goodman, Unviersity of California, Irvine

---

10:30 a.m to 12:15 p.m. — **Usability and Accessibility of Visualization Tools for Health Statistics** (Market Street Room — Invited Session) Organizers: Sue Bell and Linda Pickle, National Cancer Institute. Session Chair: Sue Bell, National Cancer Institute

10:30 a.m. — *Usability Testing of Map Designs*, Linda Williams Pickle, National Cancer Institute

10:55 a.m. — *Accessible Graphics on the World Wide Web*, Dan J. Grauman, National Cancer Institute

11:20 a.m. — *Java-based Dynamic Linked Micromap Plots*, Jim X. Chen, Xusheng Wang, and Daniel B. Carr, George Mason University, and B. Sue Bell and Linda W. Pickle, National Cancer Institute

11:45 a.m. — *Web Design and Usability Guidelines: An Evidence-based Approach*, Sanjay Koyani, National Cancer Institute

10:30 a.m to 12:15 p.m. — **Statistical Graphs** (Market West Room — Contributed Session) Session Chair: Carey Priebe, Johns Hopkins University

10:30 a.m. — *Graph Theoretic Latent Class Discovery and It's Robustness to Minimal Dominating Set Choice*, Jeff Solka, NSWCDD, Carey Priebe, JHU, and David Marchette, NSWCDD

10:50 a.m. — *Limit Theory for One-Dimensional Random Class Cover Catch Digraphs*, John Wierman and Pengfei Xiang, Johns Hopkins University

11:10 a.m. — *Bayesian Models for Sparse Edge Weighted Directed Graphs*, Deepak Agarwal, AT&T Labs

11:30 a.m. — *Sensitivity of Graph Distance Measures*, Julian Sorensen, Peter Dickinson, and Martin Schubert, Defence Science Technology Organization

11:50 a.m. — *Latent Variable Models for Link Analysis of Similarity Data.*, Juan Lin, Rutgers University

1:45 p.m to 3:30 p.m. — **Financial Risk and Fraud Detection** (Weights & Measures Room — Invited Session) Organizer and Session Chair: Scott Grimshaw, Brigham Young University

1:45 p.m. — *Computer-Aided Fraud Detection*, Conan C. Albrecht, Brigham Young University

2:10 p.m. — *Credit Scoring Using Bureau and Other Data*, Garry K. Ottosen, Millcreek Bank

2:35 p.m. — *Efficient and Secure Statistical Computing in Office Applications*, Gökhan Aydinli and Wolfgang Härdle, CASE, Humboldt-Universität zu Berlin, and Erich Neuwirth, Universität Wien

3:00 p.m. — *Some Loans are More Equal than Others: Third-party Originations and Defaults in the Subprime Mortgage Industry* , Scott D. Grimshaw, and Grant R. McQueen, Brigham Young University, William P. Alexander, Wachovia Securities, and Barrett A. Slade, Brigham Young University

1:45 p.m to 3:30 p.m. — **Computationally Challenging Statistical Methods in Genetics** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: Natascha Vukasinovic, Monsanto Company

1:45 p.m. — *Enumeration and Simulation of Marriage Node Graphs on Zero Loop Pedigrees*, Alun Thomas, University of Utah, and Chris Cannings, University of Sheffield

2:20 p.m. — *Statistical and Computational Issues in Mapping Genes in Animal Populations*, Fengxing Du, Monsanto Company

2:55 p.m. — *Computations in Animal Breeding*, Ignacy Misztal and Romdhane Rekaya, University of Georgia

---

1:45 p.m to 3:30 p.m. — **Statistical Analysis and Probabilistic Modelling of Internet Traffic** (Granary Room — Invited Session) Organizer and Session Chair: Steve Marron, University of North Carolina

1:45 p.m. — *Statistical Clustering of Internet Communication Patterns*, Felix Hernandez–Campos, A. B. Nobel, F. D. Smith, and K. Jeffay, University of North Carolina at Chapel Hill

2:20 p.m. — *A Solution to the Bandwidth Allocation Problem for the Internet*, William S. Cleveland and Jin Cao, Bell Labs, and Don X. Sun, Deephaven Capital Management

2:55 p.m. — *The Joint Distribution of Internet Flow Sizes and Durations*, Cheolwoo Park and J. S. Marron, University of North Carolina

---

1:45 p.m to 3:30 p.m. — **Data Analysis and Visualization of Graph Data** (Market Street Room — Invited Session) Organizer and Session Chair: Deborah Swayne, AT&T

1:45 p.m. — *Using Graphs to Explore Communication Networks*, Chris Volinsky, AT&T Labs-Research

2:20 p.m. — *Graphs and EDA in Computational Biology*, Robert Gentleman, Harvard University

2:55 p.m. — *Visual Exploration of Graph Data*, Deborah F. Swayne, AT&T Labs – Research, Duncan Temple Lang, Lucent Bell Laboratories, and Andreas Buja, The Wharton School, University of Pennsylvania

---

1:45 p.m to 3:30 p.m. — **Computational Methods** (Market West Room — Contributed Session) Session Chair: George Terrell, Virginia Polytechnic Institute

1:45 p.m. — *Cost-Sensitive Classifier Selection Using the ROC Convex Hull Method*, Ross Bettinger, SAS Institute

2:05 p.m. — *A New Imagery Classification Method Using Spatial Covariance Information*, James A. Shine and Daniel B. Carr, George Mason University

2:25 p.m. — *A Risk-Utility Framework for Data Swapping*, Shanti Gomatam, Alan Karr, and Ashish Sanil, National Institute of Statistical Sciences

2:45 p.m. — *On Optimal Stopping of Risk Reserve Process*, Bogdan Muciek, Wroclaw University of Technology

3:05 p.m. — *A Likelihood Approach for Determining Cluster Number*, Bill Shannon, Tsvika Klein, and Robert Culverhouse, Washington University in St. Louis School of Medicine

4:00 p.m to 5:45 p.m. — **Advances in Face/Pattern Recognition** (Weights & Measures Room — Invited Session) Organizer and Session Chair: Shailaja Suryawanshi, Merck

4:00 p.m. — *Obtaining Smooth Directional Field Estimates for Fingerprint Images*, Sarat C. Dass, Michigan State University

4:45 p.m. — *A Boosted CCCD Classifier for Fast Face Detection*, Diego Socolinsky and Joshua Neuheisel, Equinox Corporation, and Carey Priebe and Jason DeVinney, Johns Hopkins University

5:30 p.m. — Floor discussion

---

4:00 p.m to 5:45 p.m. — **Computer-Intensive Statistical Methods** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: Soumendra N. Lahiri, Iowa State University

4:00 p.m. — *Parametric Bootstrap Confidence Intervals in Small Area Estimation Problems*, Snigdhansu Chatterjee, U. Minnesota, and P. Lahiri, U. Maryland

4:25 p.m. — *Nonparametric Tail Index Estimation*, Tucker McElroy and Dimitris Politis, UC San Diego

4:50 p.m. — *Bayesian Inference in Single-Layer Neural Networks*, Robert Paige, Texas Tech University

5:15 p.m. — *A "Nonparametric Plug-in Rule" for Selecting Optimal Block Lengths for Block Bootstrap Methods*, Soumendra N. Lahiri, Iowa State University

---

4:00 p.m to 5:45 p.m. — **Computational Statistics and NSA** (Granary Room — Invited Session) Organizer and Session Chair: William Szewczyk, National Security Agency

4:00 p.m. — *Statistical Analysis of Massive Data Streams: Overview of a CATS Workshop*, Sallie Keller–McNulty, Los Alamos National Laboratory

4:35 p.m. — *A-Family Priors: Smoothing Multinomial Data*, Jeff Benedict, NSA

5:10 p.m. — *Model Building and Diagnostics for Massive Data Sets*, David W. Scott, Rice University

---

4:00 p.m to 5:45 p.m. — **Graphics and Visualization** (Market Street Room — Contributed Session) Session Chair: Scott Vander Wiel, Bell Labs, Lucent Technologies

4:00 p.m. — *Visual Data Mining for Quantized Spatial Data*, Amy Braverman, Jet Propulsion Laboratory

4:20 p.m. — *Interactive Federal Statistical Data on the Web Using "nViZn"*, Jon Hurst and Jürgen Symanzik, Utah State University

4:40 p.m. — *Visualizing Random Forests*, Adele Cutler, Utah State University, and Leo Breiman, UC Berkeley

5:00 p.m. — *Interactive Spinograms*, Heike Hofmann and Martin Theus, Iowa State University

5:20 p.m. — *Scatterplots for Massive Datasets*, Martin Theus, Di Cook, and Heike Hofmann, Iowa State University

7:00 p.m. — **Banquet** (Three Seasons Room)

**Friday, March 14, 2003**

7:30 a.m. — **Registration Continues** (Harvest Room)

8:00 a.m to 9:45 a.m. — **Sensors for Biological Threats** (Weights & Measures Room — Invited Session) Organizers: Sandra Thompson and Sharon Wunschel, Pacific Northwest National Laboratory. Chair: Sandra Thompson, Pacific Northwest National Laboratory.

8:00 a.m. — *Identification of Bio-warfare Agents and Other Applications of Molecular Biology*, Todd Ritter, Idaho Technology

8:35 a.m. — *Rapid Microbe Detection with Fluidized Bed Capture and Concentration*, Bart Weimer and Marie Walsh, Utah State University

9:10 a.m. — *Early Diagnosis of Biological Threats: Progress and Challenges*, Stephen S. Morse, Columbia University

8:00 a.m to 9:45 a.m. — **Best of KDD-2002** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: Arnold Goodman, University of California, Irvine

8:00 a.m. — *Customer Lifetime Value Modeling and Its Use for Retention Planning*, Saharon Rosset, Stanford University and Amdocs Ltd., Einat Neumann, Uri Eick, and Nurit Vatnik, Amdocs (Israel) Ltd.

8:35 a.m. — *Doing Something Useless Slightly Faster: The State of the Art in Time Series Data Mining?*, Eamonn Keogh, UC Riverside

9:10 a.m. — *Query, Analysis, and Visualization of Hierarchically Structured Data*, Christopher Stolte, Stanford University

8:00 a.m to 9:45 a.m. — **Homeland Security and Related Issues** (Granary Room — Invited Session) Organizer and Session Chair: Alan Karr, National Institute of Statistical Sciences

8:00 a.m. — *Areas of Homeland Security: At the Computational Statistical Interface*, Deborah Leishman, Los Alamos National Laboratory

8:35 a.m. — *Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks*, Andrew Moore, Carnegie Mellon University

9:10 a.m. — *Pointers from Research on Data Confidentiality and Data Quality*, Ashish Sanil, National Institute of Statistical Sciences

---

8:00 a.m to 9:45 a.m. — **Modern Text Processing, Management, and Distribution** (Market Street Room — Invited Session) Organizer and Session Chair: Jürgen Symanzik, Utah State University

8:00 a.m. — *The Journal of Statistical Software*, Jan de Leeuw, UCLA

8:35 a.m. — *Preparing Electronic Books*, Edward J. Wegman, George Mason University, and Amy Braverman, JPL

9:10 a.m. — *Electronic Books for Experts, Dummies, and Users.*, Zdenek Hlavka, Humboldt-Universität zu Berlin

---

8:00 a.m to 9:45 a.m. — **Environmental Statistics** (Market West Room — Contributed Session) Session Chair: Richard Cutler, Utah State University

8:00 a.m. — *A Spatial Model for Chronic Wasting Disease in Rocky Mountain Mule Deer*, Christopher H. Mehl and Craig J. Johns, University of Colorado at Denver

8:20 a.m. — *Spatial Statistics in the Presence of Location Error*, John Kornak and Noel Cressie, The Ohio State University

8:40 a.m. — *Separating Signal from Noise in Global Warming*, Bert W. Rust, National Institute of Standards and Technology

9:00 a.m. — *Predictive Mapping of Forest Characteristics for Fire Risk Assessment*, Gretchen Moisen and Tracey Frescino, US Forest Service, Cheng Huang and Jim Vogelmann, SAIC, US Geological Survey, and Zhiliang Zhu, US Geological Survey

9:20 a.m. — *Nonparametric Modeling of Soil Characteristics for Crop Models*, Stephan R. Sain, University of Colorado at Denver, and Doug Nychka, NCAR

---

10:30 a.m to 12:15 p.m. — **Social Networks and Statistics** (Weights & Measures Room — Invited Session) Organizer and Session Chair: Jeff Solka, Naval Surface Warfare Center

10:30 a.m. — *Random-Effects Models for Network Dependence*, Peter Hoff, University of Washington

11:05 a.m. — *Ultra-Robust and Scalable Networks Based on Hierarchies*, Peter Dodds, Duncan Watts, and Charles Sabel, Columbia University

11:40 a.m. — *Statistical Models, Degeneracy and Inference for Social Networks*, Mark S. Handcock, University of Washington

10:30 a.m to 12:15 p.m. — **Best of the International Association of Statistical Computing** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer: John Hinde, National University of Ireland, Galway. Session Chair: Wilfried Grossmann, University of Vienna

10:30 a.m. — *Incremental Algorithms for Missing Data Imputation Based on Recursive Partitioning*, Claudio Conversano, University of Cassino

11:05 a.m. — *Many Faces of a Tree*, Simon Urbanek, University of Augsburg

11:40 a.m. — *WiSP: An R Library for Simulating Wildlife Surveys*, Walter Zucchini, University of Goettingen, David Borchers, University of St. Andrews, and Stefan Kirchfeld and Martin Erdelmeier, University of Goettingen

---

10:30 a.m to 12:15 p.m. — **Digital Government Research in Support of Federal Statistics** (Granary Room — Invited Session) Organizers: Cathryn Dippo and Steve Cohen, Bureau of Labor Statistics. Session Chair: Steve Cohen, Bureau of Labor Statistics

10:30 a.m. — *Using an Ontology as Generalized Metadata Schema for Access to Distributed Heterogeneous Data Sources*, Edward Hovy, University of Southern California

11:05 a.m. — *Interfaces to a Statistical Knowledge Network*, Gary Marchionini, University of North Carolina at Chapel Hill

11:40 a.m. — *New Approaches to Mobile Computing for Field Data Collection*, Sarah Nusser, Iowa State University

---

10:30 a.m to 12:15 p.m. — **Smoothing and Nonparametric Feature Detection** (Market Street Room — Invited Session) Organizer and Session Chair: Michael Minnotte, Utah State University

10:30 a.m. — *A SiZer Analysis of IP Flow Start Times*, J. S. Marron, Felix Hernandez–Campos, and F. D. Smith, University of North Carolina

11:05 a.m. — *Longitudinal Kernel Regression*, Naisyin Wang and Raymond J. Carroll, Texas A&M University, Xihong Lin, U. Of Michigan, and Ziding Fend, Fred Hutchinson Cancer Center

11:40 a.m. — *Semiparametric Regression Smoothing and Feature Detection*, Michael G. Schimek, Karl-Franzens-University Graz, Austria

---

10:30 a.m to 12:15 p.m. — **Safety and Security** (Market West Room — Contributed Session) Session Chair: Karen Kafadar, University of Colorado at Denver

10:30 a.m. — *An Economic Index for Evaluating Traffic Safety*, Michael Conerly, J. Michael Hardin, Wade Watkins, Chunyao Feng, and Bo Hong, University of Alabama

10:50 a.m. — *Bayesian Inductively Learned Modules for Safety Critical Systems*, Jonathan E. Fieldsend, Trevor C. Bailey, Richard M. Everson, Wajtek J. Krzanowski, Derek Partridge, and Vitaly Schetinin, University of Exeter

11:10 a.m. — *Waypoint Analysis for Command and Control*, Mark Irwin, Ohio State University, David Wendt, Battelle Memorial Institute, and Noel Cressie, Ohio State University

11:30 a.m. — *Continually Improving Stream Analysis for Network Security*, Nancy J. McMillan, Douglas D. Mooney, and Dave A. Burgoon, Battelle

11:50 a.m. — *A Micro-scale Epidemiological Simulation for Management of Disease Outbreaks*, Sid Baccam, Stephen Eubank, and Catherine Macken, Los Alamos National Laboratory

---

1:45 p.m to 3:30 p.m. — **Public Health Preparedness and Response in Crisis** (Weights & Measures Room — Invited Session) Organizer and Session Chair: Jimmie D. Givens, National Center for Health Statistics

1:45 p.m. — *Using Design-Based Adaptive Sampling Procedures in Site Decontamination*, Myron J. Katzoff, Abera Wouhib, and Joe Fred Gonzalez, Jr., National Center for Health Statistics

2:25 p.m. — *Game Theory and Risk Analysis for the Smallpox Threat*, David Banks, U.S. Food and Drug Administration

3:05 p.m. — Discussant: Jimmie D. Givens, National Center for Health Statistics

---

1:45 p.m to 3:30 p.m. — **Best of the *Journal of Computational and Graphical Statistics*** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: David Scott, Rice University

1:45 p.m. — *Penalized Survival Models and Frailty*, V. Shane Pankratz, Mayo Clinic, Patricia M. Grambsch, University of Minnesota, and Terry M. Therneau, Mayo Clinic

2:20 p.m. — *Adaptive Order Selection for Spline Smoothing*, Randy Eubank, Texas A&M University, Chunfeng Huang, North Dakota State University, and Suojin Wang, Texas A&M University

2:55 p.m. — *An Adaptive Spatial Scan Density Estimation Method*, Ramani S. Pilla, Case Western Reserve University, Peng Tao, Accu Image Diagnostics Corporation, and Carey Priebe, Johns Hopkins University

---

1:45 p.m to 3:30 p.m. — **Infrastructure Security** (Granary Room — Invited Session) Organizers: Dale Anderson and Sandra Thompson, Pacific Northwest National Laboratory. Session Chair: Sandra Thompson, Pacific Northwest National Laboratory

1:45 p.m. — *Energy Infrastructure Vulnerability Assessments*, Jeff Dagle, PNNL

2:20 p.m. — *PNNL and International Border Security*, William C. Cliff, PNNL

2:55 p.m. — *Electricity Infrastructure Security*, Thomas Kropp, EPRI

---

1:45 p.m to 3:30 p.m. — **Statistical Computing** (Market Street Room — Refereed Contributed Session) Session Chair: Tim Hesterberg, Insightful Corporation

1:45 p.m. — *RGL: A R-Library for 3D Visualization with OpenGL*, Oleg Nenadic, Daniel Adler, and Walter Zucchini, University of Goettingen

2:10 p.m. — *Maximum Entropy Constructive Ensembles for Time Series Analysis*, H. D. Vinod, Fordham University

2:35 p.m. — *The Quickest Sequential Detection of Intrusions in Computer Networks*, Boris Rozovskii and Rudolf Blazek, University of Southern California, Hongjoong Kim, University of North Carolina at Charlotte, and Alexander Tartakovsky, University of Southern California

3:00 p.m. — *Implementing Legacy Statistical Algorithms in a Spreadsheet Environment*, Stephen W. Liddle and John S. Lawson, Brigham Young University

---

1:45 p.m to 3:30 p.m. — **Bioinformatics** (Market West Room — Contributed Session) Session Chair: Bart Weimer, Utah State University

1:45 p.m. — *A Bayesian Mixture Model for Bayesian Gene Expression*, Kim–Anh Do, Peter Mueller, and Feng Tang, U.T. M.D.Anderson Cancer Center

2:05 p.m. — *A Simple Approach to Accomodating Interactive and Batch Processes on a Bioinformatics Cluster*, Warren M. Snelling, John W. Keele, and Gregory P. Harhay, USDA-ARS-USMARC

2:25 p.m. — *Selecting an Optimal Rejection Region for Multiple Testing: A Decision-Theoretic Alternative to FDR Control, with an Application to Microarrays*, David R. Bickel, Medical College of Georgia

2:45 p.m. — *Statistical Methods for Spot Detection with Macroarray Data*, Adele Cutler, Andrejus Parfionovas, and Bart Weimer, Utah State University, and Yi Xie, Johns Hopkins University

---

4:00 p.m to 5:45 p.m. — **Statistical Issues in Computer Security** (Weights & Measures Room — Invited Session) Organizer and Session Chair: David Marchette, Naval Surface Warfare Center

- 4:00 p.m. — *Worm Propagation on Graphs with Heavy-tailed Degree Distribution*, Stephan Bohacek, University of Delaware
- 4:25 p.m. — *User Profiling for Intrusion Detection in Windows NT*, Tom Goldring, U.S. DOD
- 4:50 p.m. — *A Stochastic Model of Computer Intrusions for Evaluation and Exercises*, Robert P. Goldman, SIFT, LLC
- 5:15 p.m. — *Multi-Level Monitoring and Fuzzy Clustering to Detect Cyber Attacks*, Dipankar Dasgupta, Jonatan Gomez, and Fabio Gonzalez, The University of Memphis

---

4:00 p.m to 5:45 p.m. — **Interactive GeoGraphics for the Web** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: Robert M. Edsall, Arizona State University

- 4:00 p.m. — *Integrated Climate Database*, Dan Dansereau and Robert R. Gillies, Utah State University
- 4:45 p.m. — *Web Cartography for Municipal Government: An Accessibility Case Study*, Robert M. Edsall, Arizona State University
- 5:30 p.m. — Floor discussion

---

4:00 p.m to 5:45 p.m. — **Design and Statistical Computation** (Granary Room — Contributed Session) Session Chair: Bill Shannon, Washington University in St. Louis

- 4:00 p.m. — *Application of Simulated Annealing to D-optimal Design for Polynomial Regression with Correlated Observations*, Zewen Zhu and Daniel C. Coster, Utah State University
- 4:20 p.m. — *Design Aspects for Body Image Measurements*, Craig Johns, Russel Boice, and Rick Gardner, University of Colorado at Denver
- 4:40 p.m. — *A Stabilized Lugannani-Rice Formula*, George Terrell, Virginia Polytechnic Institute
- 5:00 p.m. — *Simulation from a Normally Weighted Dirichlet Distribution*, Alan Genz, Washington State University

---

4:00 p.m to 5:45 p.m. — **Nonparametrics** (Market Street Room — Refereed Contributed Session) Session Chair: Randy Eubank, Texas A&M University

- 4:00 p.m. — *Estimating Partially Linear Models Using Wavelets: A Nonlinear Backing Algorithm*, Leming Qu, Boise State University

4:25 p.m. — *A Two-Dimensional Robust Nonlinear Smoother for Irregularly Spaced Data*, Karen Kafadar, University of Colorado at Denver, and Max Morris, Iowa State University

4:50 p.m. — *A Comparison of Filter and Wrapper Methods for Feature Selection in Supervised Classification*, Edgar Acuna, University of Puerto Rico at Mayaguez

5:15 p.m. — *Novel Methods for Multivariate Ordinal Data applied to Genetic Haplotypes, Genomic Pathways, Risk Profiles, and Pattern Similarity*, Knut M. Wittkowski, The Rockefeller University

## Saturday, March 15, 2003

7:30 a.m. — **Registration Continues** (Harvest Room)

8:00 a.m to 9:45 a.m. — **Data Management for Statistical Data Bases** (Weights & Measures Room — Invited Session) Organizer and Session Chair: Andrew Westlake, Survey & Statistical Computing

8:00 a.m. — *Database Technology for Statistical Data*, Arie Shoshani, Lawrence Berkeley Laboratory

8:35 a.m. — *Metadata Usage in Statistical Computing*, Wilfried Grossmann, University of Vienna

9:10 a.m. — *Data Structures for HIV and AIDS Notification and Analysis*, Andrew Westlake, Survey & Statistical Computing

8:00 a.m to 9:45 a.m. — **Surveillance** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: Edward J. Wegman, George Mason University

8:00 a.m. — *Multiscale 'Spatial' Analysis of Network Data: Putting Wavelets on Graphs*, Eric D. Kolaczyk, Boston University

8:35 a.m. — *Social Networks and Computer Networks*, John Rigsby and Jeff Solka, Naval Surface Warfare Center Dahlgren Division

9:10 a.m. — *Classification Complexity Measures and Their Relationship to Feature Selection* , Jeff Solka and David Johannsen, Naval Surface Warfare Center Dahlgren Division

8:00 a.m to 9:45 a.m. — **Prediction of Catastrophic Events** (Granary Room — Invited Session) Organizer and Session Chair: Amy Braverman, Jet Propulsion Laboratory

8:00 a.m. — *Detecting Features in Seismic and Geodetic Data*, Andrea Donnellan and Robert Granat, Jet Propulsion Laboratory, and John Rundle, University of California, Davis

8:35 a.m. — *Predicting Damaging Climate Events: Methods, Examples, and Public Reaction*, David W. Pierce and Tim P. Barnett, Scripps Institute of Oceanography

9:10 a.m. — *Predicting and Comprehending Asteroid Impacts*, Clark R. Chapman, Southwest Research Institute

---

8:00 a.m to 9:45 a.m. — **Graphics for Bio and Chem Informatics** (Market Street Room — Invited Session) Organizer and Session Chair: Daniel B. Carr, George Mason University

8:00 a.m. — *Grapic-Centric, Computationally-Efficient Recursive Partitioning*, James Vivian, Golden Helix, Inc., S. Stan Young, CGStat LLC, and Christophe Lambert, Golden Helix, Inc.

8:45 a.m. — *Applications of Computational Geometry, Statistical Analysis, and Graphics to the Study of Molecular Systems*, Daniel B. Carr and Iosif Vaisman, George Mason University

---

10:30 a.m to 12:15 p.m. — **Statistical Methods to Compress and Query Large Databases and Data Streams** (Weights & Measures Room — Invited Session) Organizer and Session Chair: Silvia Nittel, University of Maine

10:30 a.m. — *Data Stream Algorithmics*, S. Muthukrishnan, AT&T and Rutgers University

11:05 a.m. — *Efficient Processing of Massive Data Streams for Mining and Monitoring*, Mirek Riedewald, Johannes Gehrke, Alan Demers, Abhinandan Das, and Alin Dobra, Cornell University

11:40 a.m. — *Wavelets for Efficient Querying of Large Multidimensional Data Sets*, Cyrus Shahabi, USC

---

10:30 a.m to 12:15 p.m. — **Data Mining Combat Simulations** (Smokehouse/Rap of the Gavel Room — Invited Session) Organizer and Session Chair: Barry A. Bodt, U.S. Army Research Laboratory

10:30 a.m. — *Data Mining Combat Simulations: an Emerging Opportunity*, Barry A Bodt, U.S. Army Research Laboratory

10:55 a.m. — *Regression Tree Analysis of Battle Simulation Data*, Wei–Yin Loh, University of Wisconsin, Madison

11:20 a.m. — *Robust Modeling Based on L2E Applied to Combat Simulation Data*, David Kim, United States Military Academy

11:45 a.m. — *Discovery of Battle States Knowledge from Muti-Dimensional Time Series Data*, T.W. Liao, Louisianna State University, and B. Bodt, J. Forester, C. Hansen, E. Heilman, C. Kaste, and J. O'May, U.S. Army Research Laboratory.

---

10:30 a.m to 12:15 p.m. — **Forests at Risk** (Granary Room — Invited Session) Organizer and Session Chair: Gretchen Moisen, USDA Forest Service

10:30 a.m. — *Satellite to the Public in Near Real-Time: Providing Active Wildfire Information with MODIS Rapid Response*, Mark Finco, RedCastle Resources / USDA Forest Service, Brad Quayle, USDA Forest Service, Rob Sohlberg, University of Maryland, and Jacques Descloitres, Goddard Space Flight Center

11:05 a.m. — *Identifying "Redtops": Classification of Satellite Imagery for Tracking Mountain Pine Beetle Progression through a Pine Forest*, Richard Cutler and James Powell, Utah State University, and Leslie Brown and Barbara Bentz, USDA FS Rocky Mountain Research Station

11:40 a.m. — *Design Attributes for Sampling Rare Ecological Events in Forest Ecosystems: Lichens in the Pacific Northwest*, Thomas C. Edwards and Richard Cutler, Utah State University

---

10:30 a.m to 12:15 p.m. — **Nonparametric Methods and Applications** (Market Street Room — Contributed Session) Session Chair: Zewen Zhu, Utah State University

10:30 a.m. — *Multivariate Density Estimation with Permuted Variable-Values*, Sridevi Parise, Padhraic Smyth, and Sergey Kirshner, University of California, Irvine

10:50 a.m. — *Computational Challenges in Computing Nearest Neighbor Estimates of Entropy for Large Molecules*, E. James Harner, West Virginia University, Harshinder Singh and Shengqiao Li, NIOSH/West Virginia University, and Jun Tan, West Virginia University

11:10 a.m. — *Mixture Transitions for Edge Preservation in Kalman Filtering*, Mark Fitzgerald, University of Colorado, Denver

11:30 a.m. — *Statistical Learning Theory and Statistics: Embracing New Technologies*, Kevin Watanabe, Kxen, Inc.

11:50 a.m. — *Using a LOESS Smoother to Estimate the Parameters of an Angular Dependent Distribution of HRR Data*, Bradley C. Wallet, Robert W. Hawley, and Troy L. Klein, Mission Research Corporation

---

# Interface 2003
# Proceedings and Manuscript Preparation

### Instructions for Authors submitting to the Interface Symposium and its Proceedings: Computing Science and Statistics

Beginning in the year 2000, the proceedings have been produced on a CD-ROM. This requires submission in the form of an electronic file. The ultimate production will be in the form of an Adobe Acrobat pdf file. If authors produce their own Adobe Acrobat files, they should produce a file that contains embedded fonts if at all possible. This gives much better screen resolution. However, MS Word, LaTeX, EXP, and postscript files are also acceptable. In the rare event that an author cannot provide an electronic file, we can process a paper copy with a scanner. However, results are far less satisfactory.

1. Contributed papers should not exceed 15 pages and invited papers should not exceed 20 pages. These pages include everything — bibliography, tables, graphics, title, etc.

2. Center your title horizontally on the middle five inches of the first page, and list the authors and addresses under the title. Vertically, the title should start on the first line of the page. If you cite support, do so in a footnote. Keywords and phrases may be added, but are not required.

3. Experience has shown that electronic papers (an oxymoron) appear better in a single column format. The basic text should be five inches wide horizontally and eight and 1/2 inches tall vertically. If using standard 8.5 by 11 paper, this means the side margins are one and 3/4 inches to the left of the column and one and 3/4 inches to the right of the column. The typing area for Text should begin 1 and 1/2 inches from the top of the page and leave one inch at the bottom of the page. If possible, fully justify on both left and right sides. This format is essentially the same as that found in the *Journal of Computational and Graphical Statistics*, *Computational Statistics and Data Analysis* or the *Annals of Statistics.*

4. To ensure uniformity of appearance, it is recommended to use 10 point type, preferably computer modern or Times Roman.

5. The manuscript should be single spaced. If you are familiar with TeX, this means a baseline skip of 12 points.

6. The abstract should follow the title and affiliations.

7. Other formatting is left to the discretion of the author(s).

8. The Manuscript may be submitted electronically if in TeX, LaTeX, MS Word, EXP, text, or PostScript form.

**Submission of other materials:** Authors are encouraged to submit additional supporting materials, including PowerPoint presentations (.ppt files), animations (.mpeg, .mpg, .avi, .mov files), color illustrations (.gif for line drawings, .jpg or .jpeg for continuous tone images, .tiff files), software code and datasets. Software and data should be sent in plain text files for easiest use although MS Excel or MS Access files are also satisfactory for data sets.

All material should be submitted by July 1, 2003, to `iface03@math.usu.edu`.

# Abstracts

## Short Courses

*Wednesday, 8:00 am, Weights & Measures Room:*
**Short Course I: Statistical Methods in Computer Security**

David Marchette, (NSWC), marchettedj@nswc.navy.mil

Computer security is an important aspect of homeland defense. Statistics has quite a lot to offer to the problems of detecting, modeling and defeating attacks against computers and the Internet. This course will provide a brief introduction to the network protocols, at a level sufficient to understand the data relevant to detecting and measuring network attacks. Statistical methods of detection will be discussed for various types of attacks. Topics will include attacks against networks such as denial of service attacks and attacks against hosts, such as buffer overflow attacks and user masquerading. Data analysis and visualization relevant to network and intrusion data will be discussed. Pointers to data collection utilities will be provided so that students will be able to apply what they learn in the course to their own systems and networks.

*Wednesday, 1:30 pm, Weights & Measures Room:*
**Short Course II: Topics in Computational Statistics with MATLAB**

Jeffrey L. Solka, (Naval Surface Warfare Center, Dahlgren Division), solkaj@nswc.navy.mil

The objectives of this short course are (1) to present an overview of various topics in computational statistics and (2) to show how these can be accomplished using MATLAB. The major emphasis in the course will be on the first objective. Thus, while knowledge of MATLAB would be useful, it is not required for this course. The areas we will discuss include Monte Carlo hypothesis testing, dimensionality reduction, exploratory data analysis, probability density estimation, model-based clustering and cross-validation. Methods and tools will be demonstrated within the context of an information retrieval problem.

## Keynote Address

*Thursday, 8:45 am, Three Seasons Room:*
**Considerations of Inspection for Homeland Security with Cross Linkages to Quality Control, Game Theory, and Stochastic Simulation**

James R. Thompson, (Rice University), thomp@stat.rice.edu

It is proposed to develop models for inspection (both of people and containers) utilizing insights from quality control. In Acceptance-Rejection Quality Control, we balance costs of sampling with those of passing bad items. In the Deming Paradigm of Statistical Process Control, we carry out sampling for the purposes of system improvement as opposed to lot validation. Homeland Security issues embody potential for both philosophies and we will be attempting both lot validation and system improvement. The classical Acceptance-Rejection paradigm, in the Homeland Security situation, can be shown readily to lead to inspection of all airline passengers and all baggage. It is extremely costly and leads to a situation where ticket prices soar and/or the state heavily subsidizes security.

In the United States, current allocations of funds for inspection require different strategies than 100% inspection. This leads to the use of covariate information concerning the inspected population so that the inspection should involve stratification. It also leads to strategies whereby the input stream of customers with poor risk profiles might be modified by political and other strategies.

In classical quality control, the system inspected is not sentient. In Homeland Security, the terrorist commanders are intelligent agents who will attempt to use information about inspection protocols to lessen the probability of discovery, both of terrorists and their baggage. Thus, we need to develop mixed strategies (in the sense of von Neuman-Morgenstern) hybrids for inspection strategies. Rather than seeking to deal with models simple enough for closed form solution, it is proposed to go rather toward realistic models whose analysis requires stochastic simulation .

Although the immediate charge here is for Homeland Security, it should be noted that quality control-game theoretic simulation models may also be created when planning inspections for weapons of mass destruction and/or their development in another country.

## R

(Invited Session)

Organizers: Robert Gentleman and Duncan Temple Lang
Session Chair: Robert Gentleman

*Thursday, 10:30 am, Weights & Measures Room:*
**Technology Transfer in Industry using an R-COM Interface**

Scott Vander Wiel, (Bell Labs, Lucent Technologies), scottyv@bell-labs.com,
John Chambers, (Bell Labs, Lucent Technologies), jmc@bell-labs.com,
Suresh Goyal, (Bell Labs, Lucent Technologies), goyal@lucent.com,
David James, (Bell Labs, Lucent Technologies), dj@bell-labs.com, and
Duncan Temple Lang, (Bell Laboratories, Lucent Technologies),
duncan@research.bell-labs.com

Transferring statistical methods in an industrial environment requires interfacing with engineers and managers who work primarily using PowerPoint, Excel and Word. Corporate applications such as Oracle for data warehousing and Business Objects for data access and reporting are also deeply embedded into key manufacturing and business processes.

We describe recent experience with integrating statistical smoothing and visualization into the work-flow of a project to improve manufacturing yields. Modern data analysis methods would go unused if they did not fit into a collaborative process where Excel workbooks form the basis of working team meetings and data is most readily available through Business Objects software.

A new R interface that supports Microsoft's Component Object Model (COM) is being used to orchestrate the process of fetching data from corporate repositories, smoothing and visualizing results using packages in R, and reporting through Excel. This approach has made it possible for good statistical methods to have a place alongside the traditional monthly yield tables that are generated by corporate 'business intelligence' software. The COM interface allows statisticians to work efficiently using the S language and drawing on the wealth of analytic methods available in R. It also allow process engineers and managers to work effectively because the analysis is provided in a form that integrates smoothly into their software environment.

*Thursday, 11:00 am, Weights & Measures Room:*
**Inter-System Interfaces for S**

Duncan Temple Lang, (Bell Laboratories, Lucent Technologies),
duncan@research.bell-labs.com

I will describe some of the different Omegahat packages for the S language that connect R with different languages and applications. These allow programmers in languages such as Java, Perl, Python and Octave to easily access statistical methodology and graphics from within their own familiar environment. Similarly, S users can access functionality from these other systems directly from within their familiar S programming model. In addition to these inter-language interfaces that use embedding, the RDCOMClient package allows us to dynamically and interactively access and control the many COM-enabled applications in Windows. I'll show some examples, and describe the model for the inter-system interfaces emphasizing the important aspect that data resides on the side of the interface in which it makes sense.

*Thursday, 11:30 am, Weights & Measures Room:*
**Sparse Linear Algebra for the R Language**

Pin Ng, (Northern Arizona University), Pin.Ng@nau.edu, and
Roger Koenker, (University of Illinois), roger@ysidro.econ.uiuc.edu

Many contemporary applications in statistics involve large sparse matrices, matrices with a high proportion of zero entries. Conventional array storage and associated basic linear algebra routines can be extremely burdensome for such matrices. However compressed storage schemes and specially designed algorithms can drammatically improve performance. In this talk, we will describe an implemenation of sparse linear algebra methods for the statistical language R. The implementation relies heavily on Saad's (1994) Sparskit package and the Cholesky factorization algorithms of Ng and Peyton (1993). Some performance aspects will be illustrated with applications to sparse linear regression problems including penalized L1 and L2 regression for smoothing problems. Sparse matrix classes are also useful in transferring data across platforms and software environments. A variety of sparse storage formats will be described and conversion methods discussed.

## Bioinformatics

(Invited Session)

Organizer and Session Chair: Adele Cutler

*Thursday, 10:30 am, Smokehouse/Rap of the Gavel Room:*
**Exploratory Detection of Differential Gene Expression**

John D. Storey, (University of California, Berkeley), storey@stat.berkeley.edu

We propose a statistical method for detecting differentially expressed genes in DNA microarray experiments that draws on ideas from hypothesis testing, machine learning, and false discovery rates. The final product of this methodology is a listing of genes in order of evidence for differential gene expression, as well as a gene-wise measure of significance called the q-value. Some simple arguments indicate that this method is the most powerful among all distribution free approaches to detecting differential gene expression. Numerical evidence indicates that this method outperforms the most widely used methods.

*Thursday, 11:05 am, Smokehouse/Rap of the Gavel Room:*
**Statistical Issues in the Design of Microarray Experiments**

Jean Yee Hwa Yang, (UCSF ), jean@biostat.ucsf.edu, and
Terry Speed, (UC Berkeley), terry@stat.berkeley.edu

Microarray experiments performed in many areas of biological sciences generate large and complex multivariate datasets. This talk addresses statistical design and analysis issues, which are essential to improve the efficiency and reliability of cDNA microarray experiments. We discuss various considerations unique to the design of cDNA microarrays, and examine how different types of replication affect design decisions. We calculate variances of two classes of estimates of differential gene expression based on log ratios of fluorescence intensities from cDNA microarray experiments: direct estimates, using measurements from the same slide, and indirect estimates, using measurements from different slides. These variances are compared and numerical estimates are obtained from a small experiment. Some qualitative and quantitative conclusions are drawn which have potential relevance to the design of cDNA microarray experiments.

*Thursday, 11:40 am, Smokehouse/Rap of the Gavel Room:*
**Exploration, Normalization, Summaries and Software of Affymetrix GeneChip Probe Level Data**

Rafael Irizarry, (Johns Hopkins), ririzarr@jhsph.edu

High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. Affymetrix GeneChip arrays are the most popular. In this technology each gene is typically represented by a set of 11-20 pairs of oligonucleotides separately referred to as probes. Typically 12,000 to 20,000 probe sets are arrayed on a silicon chip. RNA samples are prepared, labeled and hybridized to the arrays. Arrays are then scanned, and images produced and analyzed to obtain an intensity value for each probe. These intensities quantify the extent of the hybridization between the labeled target sample and the oligonucleotide probe. A final step to obtain expression measures is to summarize the probe intensities for a given gene in order to quantify the amount of the corresponding mRNA species in the sample. Using two extensive spike-in studies and a dilution study, we performed a careful assessment of the method of summarizing probe level data provided by the current version of the Affymetrix Microarray Suite (MAS 5.0). Careful exploration of probe level data led us to expression measures derived from empirically motivated statistical models and appropriate normalization. The new expression measure greatly improves the performance of the Affymetrix technology. The advantages of a new expression measure are assessed through bias, variance, sensitivity, and specificity. We will also describe the software used for these analyses. A paper describing this example can be found on the web: http://www.biostat.jhsph.edu/~ririzarr/papers

## Security & Infrastructure Protection: Interface of Computing Science and Statistics to the Rescue

(Invited Session)

Organizer and Session Chair: Arnold Goodman

*Thursday, 10:30 am, Granary Room:*
**Statistical Opportunities in Network Security**

David Marchette, (NSWC), marchettedj@nswc.navy.mil

In this talk I will describe some of the interesting data and problems that arise in network security applications. These problems often require the processing of vast amounts of data, and the data (and the problems) are constantly changing. This presents a nearly endless number of interesting statistical challenges, and I will highlight some of these, and discuss some areas that would benefit from the attention of statisticians.

*Thursday, 11:05 am, Granary Room:*
**Using Statistics to Detect and Thwart Denial of Service Attacks**

Carla Brodley, (Purdue University), brodley@ecn.purdue.edu

A recent phenomenon that challenges the availability of a network or host is the upswing of Denial-of-Service Attacks (NDoS), which seek to deny (or significantly degrade) a network service to the users of that service. Because NDoS attacks take many forms, characterizing an attack is not easy. Making matters worse, not all degradations of network performance are caused by malicious individuals; many times congestion due to a traffic surge or hardware/software failure can cause symptoms similar to an NDoS attack.

In this talk I will present ways in which statistics can help ameliorate the problems of NDoS. The first is to monitor the traffic for anomalous behavior and then determine whether these anomalous behaviors are due to an NDoS attack. Because signature techniques cannot detect new forms of attacks, and in the domain of network security, new attacks appear frequently, our focus is on anomaly detection. The second approach is designed to ensure that a signature-based Intrusion Detection System (IDS) cannot be thwarted by a bandwidth NDoS attack. Finally, I will briefly discuss how a host/network should react once it has detected an NDoS.

*Thursday, 11:40 am, Granary Room:*
**Maximizing Quality and Value in Security: Challenges for Computer Scientists, Statisticians and their Clients**

Arnold Goodman, (Unviersity of California, Irvine), agoodman@uci.edu

Security and infrastructure protection are currently a very-high national priority. They also pose challenges that must be met by computer scientists, statisticians and security experts, working together not only closely, but also intensely and very collaboratively.

Fundamental Challenge is to maximize the quality developed within the software and the hidden-value captured within the data, at each stage of the software's and the data's life.

Operational Challenge is for security protection to work almost all (not only some) of the time, and to account for any uncertainties outside (as well as inside) the security process.

Evaluation Challenge is to balance effort spent on analysis inside the process with effort spent on evaluation outside the process in the clients environment, difficult though it is.

Technology Challenge is to handle data, text and video and to progress from offline and reactive, through online in real time, to online and then re-configurable in real time.

Collaboration Challenge is for computer scientists, statisticians and clients to begin to recognize their dependence, and then to widen their focus until collaboration is possible.

## Usability and Accessibility of Visualization Tools for Health Statistics

(Invited Session)
Organizers: Sue Bell and Linda Pickle
Session Chair: Sue Bell

*Thursday, 10:30 am, Market Street Room:*
**Usability Testing of Map Designs**

Linda Williams Pickle, (National Cancer Institute), picklel@mail.nih.gov

Maps have the potential to display the geographic patterns of millions of statistical data points, something impossible using a tabular display. A poorly designed map, however, can fail to convey important underlying features in the data or can even distort their true geographic patterns. The National Center for Health Statistics and, later, the National Cancer Institute have conducted an interdisciplinary research program in an effort to improve map design and thus to communicate geographic statistics more effectively. Experiments were conducted on general map style, color schemes, the choice of map category cutpoints, legend design and the representation of reliability of the mapped statistics. We will summarize the results of this research and describe new geovisualization tools now being developed and tested for exploring and communicating patterns in cancer data.

*Thursday, 10:55 am, Market Street Room:*
**Accessible Graphics on the World Wide Web**

Dan J. Grauman, (National Cancer Institute), graumand@mail.nih.gov

Section 508 of the Rehabilitation Act requires that Federal agencies' electronic and information technology be accessible to people with disabilities. Text and tables on the Web are generally accessible to the visually impaired individual using a screen reader. A graphic image such as a picture can be made accessible with descriptive text called an Alt tag. Graphs and maps are more challenging. The Cancer Mortality Maps & Graphs Web site (http://cancer.gov/atlasplus/) addresses the graph challenge using dynamically-generated text files, which describe all components of the graph.

*Thursday, 11:20 am, Market Street Room:*

### Java-based Dynamic Linked Micromap Plots

Jim X. Chen, (George Mason University), jchen@cs.gmu.edu,
Xusheng Wang, (George Mason University), xwang1@gmu.edu,
Daniel B. Carr, (George Mason University), dcarr@galaxy.gmu.edu,
B. Sue Bell, (National Cancer Institute), bellsu@mail.nih.gov, and
Linda W. Pickle, (National Cancer Institute), picklel@mail.nih.gov

Linked Micromap plots (LM plots) constitute a new template for the display of spatially indexed statistical summaries. It can be used to visualize various complex data in many areas. This paper extends the existing work by introducing Java-based Dynamic LM plots, a set of dynamic LM visualization methods that allows readers to interactively select variables and modify the different views to help reveal relationships among the study units.

We use sample cancer statistics from the National Cancer Institute (NCI) as an application and implementation example to present the methods. The data set is not official data, but we believe that it provides an excellent test-bed for statistical visualization study. The system of interactive LM plots that we developed will allow NCI to present the cancer statistical summaries of the United States at the state and county level on the Internet. These Java-based dynamic LM plots have preserved all the key features of the original LM plots, and further allow better visualization through drill-down views, sorting, multiple levels of detail, magnified micromap, miniature overall statistical summary, confidence interval switching, and other interactive visualization methods to visualize the data and the relationships from different perspectives. We believe that the methods and implementations bring new ideas into statistical data visualization that allows more diversity, clarity, and convenience of presentation.

*Thursday, 11:45 am, Market Street Room:*

### Web Design and Usability Guidelines: An Evidence-based Approach

Sanjay Koyani, (Communication Technologies Branch/National Cancer Institute),
koyanis@mail.nih.gov

Studies show that Web site usability is a significant problem. About 60 percent of people who use the Web are unable to find the information they are looking for even though it exists on the site. And, in cases where people have a negative experience with a Web site, 40 percent of them never come back to that site. One key problem contributing to the significant number of unusable Web sites may be the wide range of conflicting, opinion-based Web design guidelines that exist in the field.

To address this problem, the Communication Technologies Branch (CTB) of the National Cancer Institutes Office of Communications is developing a comprehensive set of research-based Web design and usability guidelines to increase web design efficiency and effectiveness. By translating the latest research from a variety of cross-disciplinary sources into practical guidelines, CTB has been able to use the latest findings to implement effective Web design at the outset.

In early 2000, CTB published about 50 guidelines on design and usability on the Usability.gov Web site (http://usability.gov). The guidelines cover a wide variety of issues related to Web site design and development: use of white space, font type and size, navigation, labeling and use of logos and other branding components used on Web sites. Each guideline also has a rating scale showing the strength of research evidence.

Throughout most of 2002, CTB has been fully engaged in a major research project to expand the guidelines. The branch has been working with experts in the usability and design field to develop an additional 200 new design guidelines through a detailed peer review process. The new guidelines will reach across many disciplines, including technical communication, usability, human factors and cognitive psychology. The presenters will explain the extensive process involved in developing these new guidelines.

Overall, the presentation will provide:

1) an overview of why research-based Web design guidelines are important to the field; 2) recent research findings in Web design and data presentation; 3) a discussion about the process used to create the guidelines; and 4) opportunities for participants to provide input.

## Statistical Graphs

(Contributed Session)

Session Chair: Carey Priebe

*Thursday, 10:30 am, Market West Room:*

### Graph Theoretic Latent Class Discovery and It's Robustness to Minimal Dominating Set Choice

Jeff Solka, (NSWCDD), solkajl@nswc.navy.mil,
Carey Priebe, (JHU), cep@jhu.edu, and
David Marchette, (NSWCDD), marchettedj@nswc.navy.mil

This talk will present some of our recent work in the development of graph theoretic methods for latent class discovery during discriminant analysis. Applications to a gene expression data set will be provided along with discussions that detail the robustness of the procedure to the minimal dominating set choice.

*Thursday, 10:50 am, Market West Room:*

### Limit Theory for One-Dimensional Random Class Cover Catch Digraphs

John Wierman, (Johns Hopkins University), wierman@jhu.edu, and
Pengfei Xiang, (Johns Hopkins University), xiang@mts.jhu.edu

The study of class cover catch digraphs (CCCDs) is motivated by applications in nonparametric classification and pattern recognition. Priebe, Marchette, and Devinney have developed classifiers based on CCCDs for supervised classification, and also applied CCCDs to clustering problems. For the special case of uniformly distributed data in one dimension, Priebe, Marchette, and Devinney studied the exact distribution of the domination number of the data-based random CCCD, and Devinney and Wierman proved the Strong Law of Large Numbers (SLLN). This talk will discuss progress toward the SLLN and the Central Limit Theorem (CLT) for general data distributions in one dimension. The long-term goal of this investigation is to establish SLLN and CLT results for data in higher dimensions.

*Thursday, 11:10 am, Market West Room:*

### Bayesian Models for Sparse Edge Weighted Directed Graphs

Deepak Agarwal, (AT&T Labs ), dagarwal@research.att.com

We propose a new class of models based on Stochastic Blockmodels that provide global measures for a directed graph based on local interactions. The models we implement differ from the ones that already exist in the literature that focus on very small (20-30 nodes) unweighted graphs that are not too sparse. Our models apply to large (200-300 nodes), extremely sparse weighted graphs. The issue of sparseness is tackled by building Bayesian models that are known to be computationally intensive. The models are fitted using an E-M algorithm which has performed well so far. We illustrate our methodology by fitting the models to some subgraphs of a large telecommunications network.

*Thursday, 11:30 am, Market West Room:*
**Sensitivity of Graph Distance Measures**

Julian Sorensen, (Defence Science Technology Organization),
Julian.Sorensen@dsto.defence.gov.au,
Peter Dickinson, (DSTO) , Peter.Dickinson@dsto.defence.gov.au, and
Martin Schubert, (DSTO), Martin.Schubert@dsto.defence.gov.au

The performance management of computer networks is becoming increasingly important given the dynamic nature of traffic on these networks. There exists a number of graph similarity measures for network monitoring and abnormal change detection. It is necessary to quantify and compare the performance of these measures, against known types of abnormal network behaviour, to assess their suitability for use in a variety of network monitoring activities. We present a preliminary study of the performance of graph distance measures using simulated and real network traffic.

*Thursday, 11:50 am, Market West Room:*
**Latent Variable Models for Link Analysis of Similarity Data.**

Juan Lin, (Rutgers University), jklin@stat.rutgers.edu

There is a need for statistical models which can organize large collections of pair-wise similarity relationships between objects into meaningful clusters. Similarity data consists of non-negative quantitative measurements of similarity between object pairs. Examples include internet connectivity data and document word count data. We present various latent variable models for finding reduced rank structure in similarity data. Applications will be presented in unsupervised clustering, targeted clustering based on pre-defined cluster relationships, and graph layout using reduced rank graph approximations.

## Financial Risk and Fraud Detection

(Invited Session)

Organizer and Session Chair: Scott Grimshaw

*Thursday, 1:45 pm, Weights & Measures Room:*
**Computer-Aided Fraud Detection**

Conan C. Albrecht, (BYU), conan_albrecht@byu.edu

Developments in technology have made new pro-active fraud detection techniques possible. One approach using technology that appears to be effective in detecting fraud is the combination of deductive reasoning and technology—a method we call strategic fraud detection. This paper presents a model formalizing and describing the strategic fraud detection method and shows how the use of information systems provides effective ways to detect fraud.

The model includes the following six stages: (1) understanding the business, (2) identifying all possible frauds that could occur, (3) cataloging possible symptoms for each type of fraud, (4) using technology to gather data about symptoms, 5) analyzing and refining results, and (6) investigating identified symptoms. Two additional steps of (1) following up on suspected frauds and (2) automating fraud detection procedures are also discussed.

A case study applying the strategic method of fraud detection to find contractor-related frauds in a large oil refinery is presented. In this case, several frauds and other problems were discovered.

Finally, the paper makes a call for research into statistical algorithms and ratio analysis to be published in the fraud and accounting literature.

*Thursday, 2:10 pm, Weights & Measures Room:*

### Credit Scoring Using Bureau and Other Data

Garry K. Ottosen, (Chief Credit Officer, Millcreek Bank), garry.ottosen@millcreekbank.com

Credit Scoring is an intergral part of all modern-day consumer lending. Scores are build in the areas of origination, line management, collections, forecasting, and recovery. Data used in building these scores come from the credit bureaus, individual applications, and in-house master files. This presentation will focus on the types of data that are available to build scores, their usefulness, and some of the pitfalls in their use.

*Thursday, 2:35 pm, Weights & Measures Room:*

### Efficient and Secure Statistical Computing in Office Applications

Gökhan Aydinli, (CASE, Humboldt-Universität zu Berlin), aydinli@wiwi.hu-berlin.de, Wolfgang Härdle, (CASE, Humboldt-Universität zu Berlin), haerdle@wiwi.hu-berlin.de, and Erich Neuwirth, (Department of Statistics, Universität Wien), erich.neuwirth@univie.ac.at

"Let's not kid ourselves: the most widely used piece of software for statistics is Excel."

This quote of B.D. Ripley quite soberly describes the state of demand for statistical software nowadays. Not only students of economics, management science and related fields but particularly the industry asks for intuitive, efficient and secure software for statistical data analysis. This applies especially but not exclusively to the financial sector, which heavily relies on the ability to apply statistical methods in a distributed environment. But not for the sake of high implementation costs and the overhead of a steep learning curve.

The use of electronic spreadsheets as the primary software tool for teaching management science modeling techniques and quantitative methods in economics and finance undoubtedly played a key role in the increasing impact of quantitative lectures given in graduate programs. Researchers suggest that the ability to extract data from various sources and embed analytical decision models within larger systems are two of the most valuable skills for business students entering today's IT dominated workplace.

In this paper we will try to contribute to this evolution and furthermore want to argue in favor of spreadsheet applications as appropriate interface solution to matrix oriented statistical languages. We provide the addins MD*ReX and RExcel, two statistical environments embedded in Excel via (D)COM clients, based on the XploRe client /server architecture and on R as a numerical-statistical "methods server". We will emphasize the productivity gain available by combining the computational power of a statistical programming environment with the direct manipulation facilities available in spreadsheet programs like Excel. We also want to stimulate the discussion of securing the communication in such a client/server environment.

*Thursday, 3:00 pm, Weights & Measures Room:*

### Some Loans are More Equal than Others: Third-party Originations and Defaults in the Subprime Mortgage Industry

Scott D. Grimshaw, (Brigham Young University), grimshaw@byu.edu,
Grant R. McQueen, (BYU Marriott School of Management), mcqueen@byu.edu,
William P. Alexander, (Wachovia Securities), william.alexander@wachovia.com, and
Barrett A. Slade, (BYU Marriott School of Management), bslade@byu.edu

We show how agency problems between lenders (principals) and third-party originators (agents) imply that TPO-originated loans are more likely to default than similar retail-originated loans. The nature of the agency problem is that TPOs are compensated for writing loans, but not completely held accountable for the subsequent performance of those loans. Using a competing risks hazard model with unobserved heterogeneity, we find empirical support for the TPO/default prediction using individual fixed-rate subprime loans with first liens secured by residential real estate originated between 1 January 1996 and 31 December 1998. We find that apparently equal loans (similar ability-to-pay, option incentives, and term) can have unequal default probabilities. We also find that initially, the agency-cost risk was not priced. At first, the market did not recognize the higher channel-risk since TPO and retail loans received similar interest rates even though the TPO loans were more likely to default. We also show that this inefficiency was short-lived. As the difference in default rates became apparent, interest rates on TPO loans rose about 50 basis points above otherwise similar retail loans.

## Computationally Challenging Statistical Methods in Genetics

(Invited Session)

Organizer and Session Chair: Natascha Vukasinovic

*Thursday, 1:45 pm, Smokehouse/Rap of the Gavel Room:*

### Enumeration and Simulation of Marriage Node Graphs on Zero Loop Pedigrees

Alun Thomas, (University of Utah), alun@genepi.med.utah.edu, and
Chris Cannings, (University of Sheffield), c.cannings@sheffield.ac.uk

We present a method that for the marriage node graph of any zero loop pedigree will enumerate all possible pedigrees that share the same underlying tree structure. The enumeration method leads naturally to a scheme for simulating from a uniform distribution on such pedigrees. This is extended to simulating pedigrees for which the underlying marriage node graph is a tree of any particular size, and to the case when the number of individuals and the number of marriages can be independently specified.

*Thursday, 2:20 pm, Smokehouse/Rap of the Gavel Room:*

### Statistical and Computational Issues in Mapping Genes in Animal Populations

Fengxing Du, (Monsanto Company ), fengxing.du@monsanto.com

Coarse mapping of quantitative trait loci (QTL) in farm animals is commonly performed under simple designs (e.g., line cross and large half-sib families). Tracing inheritance of genes under these designs using linked markers over only a few generations is generally straightforward, but can only achieve coarse mapping. Robust statistical methods have been developed to analyze marker genotype and trait phenotype association under these designs. Complex pedigrees can achieve more experimental power for QTL detection and more accurate QTL parameter estimation, and the framework of analyzing marker data in general pedigrees using mixed models have been developed. However, most animal populations contain a very large number of inbreeding loops, and the estimation of identity by decent probability conditional on linked marker data, that is vital to marker analysis in complex pedigrees, is still a challenge. Currently, approximate methods that only use partial information and Markov chain Monte Carlo based genotype sampling algorithms are being developed to solve this problem. Fine scale mapping of QTL presents additional challenges: it requires a large number of recombinants in small chromosome regions to achieve it. Use of historical recombinants was therefore proposed, by modeling coancestry of a sample conditional on historical assumptions and observed marker data at current generations via coalescence based methods. Other challenges in QTL mapping include model selection (e.g., tightly linked QTL vs. one QTL), interaction between genes, and incorporation of expression profiling data.

*Thursday, 2:55 pm, Smokehouse/Rap of the Gavel Room:*

### Computations in Animal Breeding

Ignacy Misztal, (University of Georgia), ignacy@uga.edu, and
Romdhane Rekaya, (University of Georgia), rrelkaya@uga.edu

Most farm animals are selected for reproduction so that subsequent generations are more profitable. Information for selection includes field generated records such as weights measured at different ages, milk yields, or category of calving difficulty. Additional information is provided by pedigrees and recently by molecular markers. Statistical techniques, mainly related to mixed models, are used to separate genetic and environmental effects. System of equations may be very large. With animal populations of over 20 million and a few equations per animal, the total number of equations can exceed 100 million; however, the left hand side is usually very sparse. Two major types of computations are performed: estimation of variance components, which determine heritabilities and genetic and non-genetic relationships among various traits, and solving equations. Estimation of variance components is done either using likelihood based methods like REML, or Bayesian methods via Markov Chains. Finite methods used in computing involve sparse matrix factorization and inverse. Iterative methods involved mainly block SOR and Jacobi because of small memory requirements, although recently preconditioned conjugate gradient techniques are becoming more popular. Iterative strategies with a large number of equations are implemented matrix-free, where in each round of iteration; coefficients of the left hand side are recreated from the data. An entirely new set of problems in animal breeding arises from analysis of molecular data. System of equations become larger and less sparse. Another challenge facing animal breeders is the handling, mining and analysis of chip (microarray) data. Such data consist of the expression profiles of thousand of genes for potentially large animal populations.

## Statistical Analysis and Probabilistic Modelling of Internet Traffic

(Invited Session)

Organizer and Session Chair: Steve Marron

*Thursday, 1:45 pm, Granary Room:*

### Statistical Clustering of Internet Communication Patterns

Felix Hernandez–Campos, (University of North Carolina at Chapel Hill), fhernand@cs.unc.edu,
A. B. Nobel, (University of North Carolina at Chapel Hill), nobel@stat.unc.edu,
F. D. Smith, (University of North Carolina at Chapel Hill), smithfd@cs.unc.edu, and
K. Jeffay, (University of North Carolina at Chapel Hill), jeffay@cs.unc.edu

Internet traffic is a remarkably complex phenomenon that results from the concurrent exchange of data by a wide range of applications, such as web, email, newsgroups and file sharing. In this talk, we present our approach for grouping Internet connections into similar communication patterns using statistical clustering. Our methodology provides both a better understanding of the heterogeneity found in Internet traffic, and a new and flexible way of modeling and synthesizing the workload of the Internet.

*Thursday, 2:20 pm, Granary Room:*

### A Solution to the Bandwidth Allocation Problem for the Internet

William S. Cleveland, (Bell Labs), wsc@research.bell-labs.com,
Jin Cao, (Bell Labs), cao@research.bell-labs.com, and
Don X. Sun, (Deephaven Capital Management), dxsun@optonline.net

The most basic problem of Internet traffic engineering is determining the bandwidth (bits/sec), or link speed, required to carry a traffic load (bits/sec) offered to a single link and satisfy specified quality-of-service requirements for the traffic. The offered load is packets of varying sizes arriving for transmission on the link. Packets can queue up and are dropped if the queue size (in bits) is bigger than the buffer size (in bits) in which they are stored. For today's predominant traffic on the Internet, best-effort traffic, the applicable quality metrics are the queueing delay and the packet loss.

This bandwidth allocation problem, a critical issue for efficient engineering of the Internet, has received much attention in the network research literature. While important insight has been gained, the problem, in practical terms, has resisted solution due to a lack of comprehensive, valid statistical models for the packet arrivals and sizes. The required bandwidth depends on the queue-length process which, in turn, depends heavily on the statistical properties of the arrivals and sizes.

Equipped with recently developed statistical models for arrivals and sizes, we develop a solution by finding the bandwidth, b, required for a traffic load, t, subject to the requirements of a maximum queueing delay, d (sec), and a packet loss (percent of packets), w. The solution, a statistical model for b as a function of t, d, and w, is quite simple and employs some elements of the classical Erlang queueing delay formula for Poisson arrivals and exponential service times.

*Thursday, 2:55 pm, Granary Room:*
**The Joint Distribution of Internet Flow Sizes and Durations**

Cheolwoo Park, (UNC), cwpark@email.unc.edu, and
J. S. Marron, (UNC), marron@email.unc.edu

The joint population structure of sizes and duration times of Internet Flows is surprisingly rich. Different earlier analyses have provided nealy contradictory answers. This controvesy is resolved here. In particular, by considering a global family of data thresholdings, we show that the differing results are driven by different types of thresholding, before the computation of log-log correlations.

## Data Analysis and Visualization of Graph Data

(Invited Session)

Organizer and Session Chair: Deborah Swayne

*Thursday, 1:45 pm, Market Street Room:*
**Using Graphs to Explore Communication Networks**

Chris Volinsky, (AT&T Labs-Research), volinsky@research.att.com

When studying transactional networks such as telephone call detail data, credit card transactions, or web clickstream data, graphs are a convenient and informative way to represent data. When the graph edges represents actual communications between transactors, the graph can then be mined to find communities of communicators, for the purpose of detecting fraud cells or marketing segments. Through a combination of visualization, graph theory algorithms, and statistical analysis, we can learn things from the graph that we could not have discovered otherwise. In this talk I will introduce some graphs from our communications networks, and discuss how we have used these tools to find interesting communities.

*Thursday, 2:20 pm, Market Street Room:*
**Graphs and EDA in Computational Biology**

Robert Gentleman, (Dept. of Biostatistics, Harvard), rgentlem@hsph.harvard.edu

Graphs provide a unique data structure for exploring biological data. There are many different graphs that can be constructed based on biologic data. These include metabolic pathways, protein-protein interactions as well as co-citation of genes in the scientific literature. In this talk I will consider various methods of using graphs and their properties to perform exploratory data analysis (EDA) on data from a microarray experiment using different graphs based on biological meta-data.

*Thursday, 2:55 pm, Market Street Room:*
**Visual Exploration of Graph Data**

Deborah F. Swayne, (AT&T Labs – Research), dfs@research.att.com,
Duncan Temple Lang, (Lucent Bell Laboratories), duncan@research.bell-labs.com, and
Andreas Buja, (The Wharton School, U. of Pennsylvania), andreas.buja@wharton.upenn.edu

Graphs have long been of interest in telecommunications and social network analysis, and they are now receiving increasing attention from statisticians working in other areas, particularly in biostatistics.

Most of the visualization software available for working with graphs has come from outside statistics and has not included the kind of interaction that statisticians have come to expect. At the same time, most of the exploratory visualization software available to statisticians has made no provision for the special structure of graphs.

Graphics software for the exploratory visual analysis of graph data should include the following: graph layout methods; a variety of displays and methods for exploring variables on both nodes and edges, including methods that allow these covariate displays to be linked to the network view; methods for thinning a dense graph. In addition, the power of the visualization software is greater if it can be smoothly linked to an extensible and interactive statistics environment.

In this talk, we'll describe and demonstrate how these goals have been addressed in GGobi through its data format, graphical user interface design, and its relationship to the R software.

## Computational Methods

(Contributed Session)

Session Chair: George Terrell

*Thursday, 1:45 pm, Market West Room:*
**Cost-Sensitive Classifier Selection Using the ROC Convex Hull Method**

Ross Bettinger, (SAS Institute), Ross.Bettinger@sas.com

One binary classifier may be preferred to another based on the fact that it has better prediction accuracy than its competitor. Without additional information describing the cost of a misclassification, accuracy alone as a selection criterion may not be a sufficiently robust measure when the distribution of classes is greatly skewed or the costs of different types of errors may be significantly different.

The receiver operating characteristic (ROC) curve is often used to summarize binary classifier performance due to its ease of interpretation, but does not include misclassification cost information in its formulation. Provost and Fawcett have developed the ROC Convex Hull (ROCCH) method that incorporates techniques from ROC curve analysis, decision analysis, and computational geometry in the search for the optimal classifier that is robust with respect to skewed or imprecise class distributions and disparate misclassification costs.

We apply the ROCCH method to several datasets using a variety of modeling tools to build bi-nary classifiers and compare their performances using misclassification costs. We support Provost, Fawcett, and Kohavis claim that classifier accuracy, as represented by the area under the ROC curve, is not an optimal criterion in itself for choosing a classifier, and that by using the ROCCH method, a more appropriate classifier may be found that realistically reflects class distribution and misclassification costs.

*Thursday, 2:05 pm, Market West Room:*

### A New Imagery Classification Method Using Spatial Covariance Information

James A. Shine, (George Mason University), jshine1@gmu.edu, and
Daniel B. Carr, (George Mason University), dcarr@gmu.edu

Classical and modern statistical methods offer a wide variety of approaches to classification of data in general and classification of imagery in particular. None of these approaches explicitly use spatial information. Spatial covariance structures have been used for data prediction, but not directly for classification. This paper describes a classification method using the spatial covariance information in imagery to directly classify images in a supervised approach. A series of thresholds are measured with training data for each class, and a model is then fitted. Each pixel is measured for its fit for each class, and the class with the best fit is chosen. A framework is also described for using multiple bands of information and classifying from the combined bands. Results from multispectral imagery classification will be discussed and analyzed.

*Thursday, 2:25 pm, Market West Room:*

### A Risk-Utility Framework for Data Swapping

Shanti Gomatam, (National Inst. of Stat. Sciences), sgomatam@niss.org,
Alan Karr, (NISS), karr@niss.org, and
Ashish Sanil, (NISS), ashish@niss.org

Data swapping is a common statistical disclosure limitation method used to protect the confidentiality of data. We consider the problem of selecting an optimal release when data swapping is applied to categorical variables. Risk and utility values associated with every post-swap release are traded off to isolate a release "frontier" which contains the optimal release(s). Current Population Survey data are used to illustrate the technique, and results obtained when considering several swap variables and multiple swap rates are presented.

*Thursday, 2:45 pm, Market West Room:*

### On Optimal Stopping of Risk Reserve Process

Bogdan Muciek, (Institute of Mathematics, Wroclaw University of Technology, Wroclaw, Poland), B.Muciek@im.pwr.wroc.pl

The following problem in risk theory is considered. An insurance company, endowed with an initial capital, receives premiums and pays out claims that occur according to a renewal process. The times between consecutive claims are independent and identically distributed (i.i.d.). The sequence of successive claims is a sequence of i.i.d. random variables. The capital of the company is invested with an interest rate, claims increase with another rate. The aim is to find the stopping time that maximizes the capital of the company. A dynamic programming method is used to find the optimal stopping time and to specify the expected capital at that time. The cases of immidiate claim payout as well as at the end of periods are considered.

*Thursday, 3:05 pm, Market West Room:*

### A Likelihood Approach for Determining Cluster Number

Bill Shannon, (Washington Univ. in St. Louis School of Medicine), shannon@ilya.wustl.edu,
Tsvika Klein, (Washington Univ. in St. Louis School of Medicine), tklein@im.wustl.edu, and
Robert Culverhouse, (Washington Univ. in St. Louis School of Medicine), rob@frodo.wustl.edu

Deciding where to cut the dendrogram produced by a hierarchical cluster analysis is known as as the stopping rule problem. Heuristic approaches proposed for solving this problem have been based on statistics such as the proportion of variance accounted for by the clusters. Such measures are based on reasonable ad hoc measures, not on a probability model of cluster distributions. The statistic is calculated on each of the sets of clusters produced by cutting the dendrogram at successive heights. The number of clusters in the set that optimizes the statistic estimates the true number of clusters.

In this presentation we propose a novel stopping rule based on a probability model for graphical objects. The application of probability models to hierarchical trees is highly speculative, but is based on prior published work (Shannon and Banks 1999; Banks and Constantine 1999; McMorris and Major 1990). We propose to extend this prior work to derive a likelihood or likelihood-ratio test (LRT) for determining the number of clusters in a dataset. We are aware that the criteria for the LRT (Lehman 1999) are not fully met so that P values based on it will be approximations at best, though bootstrap P values might easily be estimated. We are beginning to contrast the likelihood and likelihood-ratio test stopping rule with other exsiting ad hoc approaches. In our talk we present this method for the first time and show some very preliminary results.

## Advances in Face/Pattern Recognition

(Invited Session)

Organizer and Session Chair: Shailaja Suryawanshi

*Thursday, 4:00 pm, Weights & Measures Room:*
**Obtaining Smooth Directional Field Estimates for Fingerprint Images**

Sarat C. Dass, (Michigan State University), sdass@msu.edu

Fast and robust estimation of the directional field (DF) is fundamental to the processing of fingerprint images. The estimation of the DF is approached from the point of view of Bayesian statistics. Distributional models are assumed for the observed gradients given the unknown underlying principal gradient directions. Spatial smoothness of the DF in a fingerprint image is modelled using a class of Markov random field priors. The Maximum-A-Posteriori (MAP) estimate of the DF obtained exhibits spatial smoothness while preserving important singularities in the image. We develop algorithms to compute this MAP estimate of the DF in real time. The general framework presented here encompasses previous work on DF estimation as special cases. We also present the results of the DF estimation on each fingerprint image belonging to the Henry class.

*Thursday, 4:45 pm, Weights & Measures Room:*
**A Boosted CCCD Classifier for Fast Face Detection**

Diego Socolinsky, (Equinox Corporation), diego@equinoxsensors.com,
Joshua Neuheisel, (Equinox Corporation), jneuheisel@equinoxsensors.com,
Carey Priebe, (Johns Hopkins University), cep@jhu.edu, and
Jason DeVinney, (Johns Hopkins University), devinney@mts.jhu.edu

We introduce a fast object detection algorithm based on the Class-Cover Catch Digraph (CCCD) classifier. When applied to face detection, our algorithm exhibits excellent performance with speeds close to those of the fastest reported techniques. The main technical innovations in our method include a boosted tree-like CCCD classifier with a maximum rejection bias, and the use of a cross-correlation metric for fast similarity computation.

## Computer-Intensive Statistical Methods

(Invited Session)

Organizer and Session Chair: Soumendra N. Lahiri

*Thursday, 4:00 pm, Smokehouse/Rap of the Gavel Room:*
**Parametric Bootstrap Confidence Intervals in Small Area Estimation Problems**

Snigdhansu Chatterjee, (U. Minnesota), chatterjee@stat.umn.edu, and
P. Lahiri, (U. Maryland), plahiri@survey.umd.edu

We propose a parametric bootstrap method to construct confidence interval for a small area mean in the Fay-Herriot model. The method is accurate for coverage probability up to $O(n^{-3/2})$, a property not established for competing methods. Simulations show among competing techniques it also has the shortest length.

*Thursday, 4:25 pm, Smokehouse/Rap of the Gavel Room:*
**Nonparametric Tail Index Estimation**

Tucker McElroy, (UC San Diego), tmcelroy@math.ucsd.edu, and
Dimitris Politis, (UC San Diego), politis@math.ucsd.edu

In Politis (2002) a method of tail index estimation for heavy-tailed time series, based on examining the growth rate of the logged sample second moment of the data, was proposed and studied. This estimator has a slow rate of convergence to the tail index, which is due to the high dependence of the summands of the statistic. To ameliorate the convergence rate, this work proposes an estimator with faster convergence rate and reduced bias, which is computed over subblocks of the whole data set. The resulting estimator obtains a polynomial rate of consistency for the tail index, and in simulation studies shows itself decidedly superior to competing prior art, such as the above-mentioned estimator of Politis (2002), as well as the reknowned Hill estimator. The use of blocks, which is computationally intensive, gives superior results over a wide range of heavy-tailed models, including those in the non-normal domain of attraction, which cannot be handled by the estimator of Politis (2002).

*Thursday, 4:50 pm, Smokehouse/Rap of the Gavel Room:*
**Bayesian Inference in Single-Layer Neural Networks**

Robert Paige, (Texas Tech University), rpaige@math.ttu.edu

Approximate marginal Bayesian computation and inference are developed for single-layer neural network models. In particular we consider the classical neural network which uses basis functions and a novel wavelet neural network with wavelet basis functions. The marginal considerations include determination of approximate Bayes factors for model choice, approximate predictive density computation for a future observable, and determination of approximate Bayes estimates for the nonlinear regression function. Standard conjugate analysis applied to the linear parameters leads to an explicit posterior on the nonlinear parameters. The proposed methodology is illustrated in the context of a nonlinear dataset which involves a univariate nonlinear regression model.

*Thursday, 5:15 pm, Smokehouse/Rap of the Gavel Room:*
**A 'Nonparametric Plug-In Rule' for Selecting Optimal Block Lengths for Block Bootstrap Methods**

Soumendra N. Lahiri, (Iowa State University), snlahiri@iastate.edu

In this talk, we consider the problem of choosing the optimal block size for block bootstrap methods empirically. We suggest a general plug-in principle for estimating MSE-optimal smoothing parameters, and establish its validity for estimating the of the optimal block size for block bootstrap methods. Unlike standard plug-in rules, the proposed method can be applied without explicit analytical expressions for the constants that appear in the leading term of the optimal block size. We also discuss computational efficacy of the method and illustrate its finite sample properties through numerical examples.

## Computational Statistics and NSA

(Invited Session)

Organizer and Session Chair: William Szewczyk

*Thursday, 4:00 pm, Granary Room:*

### Statistical Analysis of Massive Data Streams: Overview of a CATS Workshop

Sallie Keller–McNulty, (Los Alamos National Laboratory), sallie@lanl.gov

The National Research Council's Committee on Applied and Theoretical Statistics (CATS) recently held a two-day workshop exploring methods for the statistical analysis of streams of data so as to stimulate further progress in this field. "Data Streams" may be defined as "A sequence of digitally encoded signals used to represent information in transmission". The workshop focus was on data streams that are too massive or dynamic to be subjected to batch processing. With such data streams, massive amounts of data are arriving continually and it is necessary to perform very frequent analyses or re-analyses on the constantly arriving data. Often there is so much data that only a short time window's worth is economically storable, necessitating summarization strategies. The workshop brought together a broad base of researchers from statistics, probability, and computer science who are dealing with massive data streams in different contexts. Sessions were held on the following topics: Atmospheric and Meteorological Data; High-Energy Physics; Integrated Data Streams; Network Traffic ; and Mining Commercial Streams of Data. The workshop was very successful in starting the crossfertilization of ideas among attendees. This talk will give an overview of the workshop's activities, with a focus on recommendations made during discussion sessions ranging from exciting research problems to ideas for increased collaborations.

*Thursday, 4:35 pm, Granary Room:*

### A-Family Priors: Smoothing Multinomial Data

Jeff Benedict, (NSA),

The A-family is a family of distributions for probability vectors. In this talk, the A-family serves as a prior distribution for probability vectors that generate multinomially distributed count vectors. Posterior inference based on importance sampling is described. The A-family is used to model prior smoothness in frequency histograms and this idea is applied to stamp-thickness counts.

*Thursday, 5:10 pm, Granary Room:*

### Model Building and Diagnostics for Massive Data Sets

David W. Scott, (Rice University), scottdw@rice.edu

The process of statistical modeling is an iterative process, which culminates with model modification if indicated by failure in the diagnostics phase. This process of model modification is an art, but perhaps still poorly understood and difficult to teach. We discuss the extra difficulties in model diagnostics and modification resulting from the extraordinary burdens of working in the world of massive data sets.

## Graphics and Visualization

(Contributed Session)

Session Chair: Scott Vander Wiel

*Thursday, 4:00 pm, Market Street Room:*

### Visual Data Mining for Quantized Spatial Data

Amy Braverman, (Jet Propulsion Laboratory), Amy.Braverman@jpl.nasa.gov

We discuss a visual data mining environment for quantized, multivariate spatial data sets generated by remote sensing instruments. Typically, remote sensing data are difficult to explore on a global scale because of their size, spatial, temporal and multivariate complexity, and hierarchical structure. We previously proposed (Braverman, JCGS March 2002) quantizing such data sets as a means of reducing their size and complexity. Specifically, data are partitioned on a spatial-temporal grid (e.g. one degree latitude by longitude by month), and data in each grid cell replaced by a set of differentially weighted representative values. The weights show how many of the original data points are represented by each representative. Here, we propose applying a modified version of the same methodology to the representatives themselves to achieve subsequent levels of data and complexity reduction. This allows us to coarsen both spatial and quantization resolution, facilitating better understanding of spatial, multivariate relationships at various levels of the hierarchy. We introduce a java visualization tool that allows interactive data exploration according to this model.

*Thursday, 4:20 pm, Market Street Room:*

### Interactive Federal Statistical Data on the Web Using "nViZn"

Jon Hurst, (Utah State University), jon@jonathan.hurst.name, and
Jürgen Symanzik, (Utah State University), symanzik@math.usu.edu

Online applications are an attractive solution for providing quick access to geographically referenced federal data sets. In the past, available software was not ideally suited for interactive, statistical graphics applications on the Web. "nViZn" (read envision) is a Java-based software development kit for statistical graphics. Building on the "nViZn" libraries, we developed software for the interactive display of federal air quality data. This software allows users to display, sort, and compare multiple tables and micromaps. Having produced this display framework we conclude that "nViZn" based applications are a good solution for interactive statistical graphics on the Internet.

*Thursday, 4:40 pm, Market Street Room:*

### Visualizing Random Forests

Adele Cutler, (Utah State University), adele@stat.usu.edu, and
Leo Breiman, (UC Berkeley), leo@stat.berkeley.edu

Random forest classifiers are as accurate as support vector machines, but have the potential to be much more interpretable. We describe random forest classifiers and introduce visualization software that is being developed to help users interpret the results. The software is based on the java component library VisAD (http://www.ssec.wisc.edu/~billh/visad.html).

*Thursday, 5:00 pm, Market Street Room:*
**Interactive Spinograms**

Heike Hofmann, (Iowa State University), hofmann@iastate.edu, and
Martin Theus, (Iowa State University), martin@TheusRus.de

For one-dimensional explorations, histograms and barcharts are commonly used to visualize continuous and categorical variables, respectively.

In higher dimensions corresponding diagrams are scatterplots and mosaicplots. However, if we deal with a mixture of continuous and discrete data, the choice of a single display is not that obvious any more. Along with interactive linking, Spinograms provide a solution to this problem. As a continuous counterpart Spinograms derive from histograms as spineplots (Hummel, 1996) do from barcharts. Spinograms can be used to visualize (empirical) conditional densities, which is an almost impossible job when using box-plots or histograms.

Examples show how useful Spinograms are in practice. But they are not only easily understood and widely applicable, they also enjoy some important excellent theoretical properties: they provide a simple visual test on the distributions of subgroups compared to the whole sample; trends in the data are easy to spot and to formulate in a statistical model. Altogether, spinograms provide a powerful tool for a fast exploratory and interactive analysis.

*Thursday, 5:20 pm, Market Street Room:*
**Scatterplots for Massive Datasets**

Martin Theus, (Iowa State University), theusm@iastate.edu,
Di Cook, (Iowa State University), dicook@iastate.edu, and
Heike Hofmann, (Iowa State University), hofmann@iastate.edu

Representing data in scatterplots works well up to about tens of thousands cases. A point on a page takes very little ink so a lot of points can be drawn before overplotting occurs, especially when optimizations such as pixel sized glyphs and large plot windows are used. However scatterplots lose their usefulness when data sets reach the order of 100k. With such large data substantial overplotting masks structure in the data. Thus plots of large data are inherently binned by the screen real estate. This talk discusses the use of alpha-blending to render the points and grey scale to represent the counts at each pixel, investigating these methods for the representation of pairs of variables of large data.

Now additionally data visualization is most useful when implemented into an interactive system that allows linking information between several plots. This talk also investigates the nature of linking information from a binned sctterplot representation.

## Sensors for Biological Threats

(Invited Session)

Organizers: Sandra Thompson and Sharon Wunschel
Session Chair: Sandra Thompson

*Friday, 8:00 am, Weights & Measures Room:*
**Identification of Bio-warfare Agents and Other Applications of Molecular Biology**

Todd Ritter, (Idaho Technology), todd@idahotech.com

Techniques in molecular biology have long been used for specific, sensitive identification of microorganisms (i.e. bacteria and viruses). Military users rely on such technology for identifying safe food, water sources, natural endemic diseases (may reduce readiness), and defense against biological warfare. The evolution of these techniques has moved the science from the lab to the broad area of real world use (e.g. military). Advancements in software, hardware and chemistry allow trained technicians to produce reference laboratory quality results in some of the most austere environments. Multi-use technologies reduce the logistics load for militaries by testing multiple samples for different organisms.

For example, in 1999, US Forces in Saudi Arabia quickly identified a Salmonella outbreak by testing various clinical, food and environmental samples. The fast and accurate identification limited the spread of the outbreak to less than 3directed health care providers towards the proper antibiotics for treatment. Additionally, such applications keep force structure safe from environmental or local health threats. Using these rapid techniques helps posture military forces for deployment to remote sites. For example, a force deploying to an area with an endemic disease may choose to test local water sources, insects, and other vectors before exposing immunologically naive personnel. This allows the appropriate measures to be taken such as choosing proper vaccinations and/or antibiotics. Such information adds a new dimension to military readiness. As a result, operations are deployed with greater cost savings and improved efficiency.

Proactively testing different infectious agents provides valuable data and insight to commanders and health officials. From a military readiness perspective, a biowarfare threat is a paramount concern. Until recently, a biological attack or induced epidemic would go undetected until large numbers of patients became ill. With the Internet and other advancements in information technology, rapid diagnostics can be used as surveillance tools. The US Air Force developed a medical/bio-warfare surveillance system called L.E.A.D.E.R. (Lightweight Epidemiology and Advanced Detection, and Emergency Response System). LEADER is a medical surveillance tool providing real-time analysis of data to identify the presence of a covert or naturally occurring bio-event. Clinical data is collected using specific medical applications and laboratory identification tools (i.e. RAPID by Idaho Technology). The data is stored and analyzed at a remote central repository. An alert generates when a pattern of symptoms indicates suspect activity.

Operationally, this system presents data from areas usually not shared in a single, unified format. This enables critical decision-making data to be placed in a command suite to provide commanders informational superiority. Ultimately, this will maintain troop readiness as a force protection tool. In summary, the LEADER system is a suite of tools that provides timely diagnostics and consultative capability in a light, highly mobile platform.

*Friday, 8:35 am, Weights & Measures Room:*
**Rapid Microbe Detection with Fluidized Bed Capture and Concentration**

Bart Weimer, (Utah State University), bcweimer@cc.usu.edu, and
Marie Walsh, (Utah State University), mkwalsh@cc.usu.edu

Rapid bacterial detection is a key issue in a various applications. Many methods are available that quickly detection organisms after an enrichment or growth step. Other technologies are available for use with small sample sizes. The Center for Microbe Detection and Physiology is focused on development of technologies to eliminate growth and are useful with larger samples. Various technologies that capture and concentrate organisms onto a solid surface have been developed using a fluidized bed format with high flow rates. One such target-specific technology, ImmunoFlow, uses antibodies to capture and detect target organisms. This format successfully eliminates growth, uses 50 ml to 5000 ml samples, and is done within 30 minutes. Other target non-specific technologies capture and concentration toxins and organisms via biomimetic schemes. When coupled with PCR detection these assays provide a two-step technique to provide a presumptive and confirmed result.

*Friday, 9:10 am, Weights & Measures Room:*
**Early Diagnosis of Biological Threats: Progress and Challenges**

Stephen S. Morse, (Columbia University), ssm20@columbia.edu

The combination of recent advances in detection technology and molecular biology (such as automated real-time PCR), and in informatics and network communications, have made possible major improvements in agent detection and identification, with the potential to revolutionize both environmental detection and medical diagnostic capabilities (as discussed by Ritter). However, as demonstrated by the recent anthrax events, the first indication of a bioterrorist attack may well be the appearance in emergency rooms or doctors' offices of people sick with an unexpected illness, and the public health and medical responses may well be underway before the true nature of the event is recognized. Conceptually, many of the steps that the public health system needs to take in order to strengthen our national biodefense are very similar to what needs to be done to prepare for an unexpected naturally occurring outbreak of infectious disease (what we may call "emerging infections plus"). Early recognition and treatment are essential to save lives, but many of these diseases are nonspecific in the early stages, often just resembling the flu, and the agent may be virtually absent in normal diagnostic samples collected early in infection. A major diagnostic challenge therefore remains identifying those who are exposed, and differentiating them from the larger number of "worried well". Despite the difficulties, some progress has been made recently in this "pre-symptomatic" or "prodromal" diagnosis. Some current strategies to be discussed include identification of early host responses to infection, and detection of agent or host response markers in breath or other easily accessed noninvasive samples.

(Supported by grants from the Arts & Letters Foundation, the Achelis Foundation, and through CDC/ASPH cooperative agreement for the Centers for Public Health Preparedness.)

## Best of KDD-2002

(Invited Session)

Organizer and Session Chair: Arnold Goodman

*Friday, 8:00 am, Smokehouse/Rap of the Gavel Room:*

### Customer Lifetime Value Modeling and Its Use for Retention Planning

Saharon Rosset, (Stanford University and Amdocs Ltd.), saharon@stanford.edu,
Einat Neumann, (Amdocs (Israel) Ltd.), einatn@amdocs.com,
Uri Eick, (Amdocs (Israel) Ltd.), urieick@amdocs.com, and
Nurit Vatnik, (Amdocs (Israel) Ltd.), nuritv@amdocs.com

We present and discuss the important business problem of estimating the effect of retention efforts on the Lifetime Value of a customer in the Telecommunications industry. We discuss the components of this problem, in particular customer value and length of service (or tenure) modeling, and present a novel segment-based approach, motivated by the segment-level view marketing analysts usually employ. We then describe how we build on this approach to estimate the effects of retention on Lifetime Value. Our solution has been successfully implemented in Amdocs Business Insight (BI) platform, and we illustrate its usefulness in real-world scenarios.

*Friday, 8:35 am, Smokehouse/Rap of the Gavel Room:*

### Doing Something Useless Slightly Faster: The State of the Art in Time Series Data Mining?

Eamonn Keogh, (UC Riverside), eamonn@cs.ucr.edu

In the last decade there has been an explosion of interest in mining time series data. Literally hundreds of papers have introduced new algorithms to index, classify, cluster and segment time series. In this work we make the following claim. Much of this work has very little utility because the contribution made (speed in the case of indexing, accuracy in the case of classification and clustering, model accuracy in the case of segmentation) offer an amount of "improvement" that would have been completely dwarfed by the variance that would have been observed by testing on many real world datasets, or the variance that would have been observed by changing minor (unstated) implementation details.

To illustrate our point, we have undertaken the most exhaustive set of time series experiments ever attempted, re-implementing the contribution of more than two dozen papers, and testing them on 50 real world, highly diverse datasets. Our empirical results strongly support our assertion, and suggest the need for a set of time series benchmarks and more careful empirical evaluation in the data mining community.

*Friday, 9:10 am, Smokehouse/Rap of the Gavel Room:*
**Query, Analysis, and Visualization of Hierarchically Structured Data**

Christopher Stolte, (Stanford University), cstolte@stanford.edu

In the last several years, large OLAP databases have become common in a variety of applications such as corporate data warehouses and scientific computing. To support interactive analysis, many of these databases are augmented with hierarchical structures that provide meaningful levels of abstraction that can be leveraged by both the computer and analyst. This hierarchical structure generates many challenges and opportunities in the design of systems for the query, analysis, and visualization of these databases.

In this talk, I'll present an interactive visual exploration tool that facilitates exploratory analysis of data warehouses with rich hierarchical structure, such as might be stored in data cubes. We based this tool on Polaris, a system for rapidly constructing table-based graphical displays of multidimensional databases. Polaris builds visualizations using an algebraic formalism that is derived from the interface and interpreted as a set of queries to a database. We have extended the user interface, algebraic formalism, and generation of data queries in Polaris to expose and take advantage of hierarchical structure. In the resulting system, analysts can navigate through the hierarchical projections of a database, rapidly and incrementally generating visualizations for each projection.

## Homeland Security and Related Issues

(Invited Session)

Organizer and Session Chair: Alan Karr

*Friday, 8:00 am, Granary Room:*
**Areas of Homeland Security: At the Computational Statistical Interface**

Deborah Leishman, (Los Alamos National Laboratory), leishman@lanl.gov

This talk will outline work being done in several areas within Homeland Security and identify needs for close cooperation and integration of specific computational and statistical methods to achieve success. In particular we will cover application areas such as surveillance, terrorist tracking and image analysis and relate these to the required methods both computational and statistical that underlie them. In particular, this description will describe capabilities such as knowledge integration that are needed at the application level as well as and those required at a more detailed technical computer infrastructure level.

*Friday, 8:35 am, Granary Room:*
**Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks**

Andrew Moore, (Carnegie Mellon University), awm@cs.cmu.edu

Joint work with Weng-Keen Wong, Gregory Cooper and Michael Wagner.

We present an algorithm for performing early detection of disease outbreaks by searching a database of emergency department cases for anomalous patterns. Traditional techniques for anomaly detection are unsatisfactory for this problem because they identify individual data points that are rare due to particular combinations of features. When applied to our scenario, these traditional algorithms discover isolated outliers of particularly strange events, such as someone accidentally shooting their ear, that are not indicative of a new outbreak. Instead, we would like to detect anomalous patterns. These patterns are groups with specific characteristics whose recent pattern of illness is anomalous relative to historical patterns. We propose using a rule-based anomaly detection algorithm that characterizes each anomalous pattern with a rule. The significance of each rule is carefully evaluated using Fisher's Exact Test and a randomization test. Our algorithm is compared against a standard detection algorithm by measuring the number of false positives and the timeliness of detection. Simulated data, produced by a simulator that creates the effects of an epidemic on a city, is used for evaluation. The results indicate that our algorithm has significantly better detection times for common significance thresholds while having a slightly higher false positive rate.

*Friday, 9:10 am, Granary Room:*
**Pointers from Research on Data Confidentiality and Data Quality**

Ashish Sanil, (National Institute of Statistical Sciences), ashish@niss.org

Issues of Data Confidentiality (DC) – making information available (possibly by releasing suitably modified/restricted data sets) but at the same time protecting the confidentiality of data subjects – and Data Quality (DQ) have been enduring concerns of Federal statistical agencies and other agencies that regularly disseminate data to the public. Consequently, there exists a body of established techniques for addressing DC and DQ problems. There are dual aspects to the DC and DQ problems that are relevant to homeland security issues. In case of DC, in order to apply disclosure limitation methods to the data, one has to understand how a possibly malicious user could identify individual in the released data. A significant portion of DQ efforts involve the detection of anomalous records in the data. Both these families of techniques can be useful in unmasking individuals in databases with noisy data, and for identifying aberrant records in the database. Lessons gleaned from NISS projects on DC and DQ will be presented as illustrations.

## Modern Text Processing, Management, and Distribution

(Invited Session)

Organizer and Session Chair: Jürgen Symanzik

*Friday, 8:00 am, Market Street Room:*
**The Journal of Statistical Software**

Jan de Leeuw, (UCLA Statistics), deleeuw@stat.ucla.edu

The Journal of Statistical Software (www.jstatsoft.org) is now in its seventh year. I review the reasons for establishing the journal, evaluate its successfullness, discuss its format and the specific problems in its publication process. I also compare JSS with other electronic and paper journals, and mention possible extensions and improvements.

*Friday, 8:35 am, Market Street Room:*
**Preparing Electronic Books**

Edward J. Wegman, (George Mason University), ewegman@gmu.edu, and
Amy Braverman, (JPL), Amy.J.Braverman@jpl.nasa.gov

Interface has for several years been issuing Proceedings on CDs. The process of preparing books/proceedings for CD or other electronic publication involves a number of steps and a variety software. This paper will discuss the preparation process and some of the subtleties involved.

*Friday, 9:10 am, Market Street Room:*
**Electronic Books for Experts, Dummies, and Users.**

Zdenek Hlavka, (Humboldt-Universität zu Berlin), hlavka@wiwi.hu-berlin.de

The beauty of electronic books lies in the wide range of formats which allow different views of just one text. The user chooses freely the format which best suits his needs. An example is, e.g., a choice between the PDF and HTML version: PDF typically features much better typesetting whereas viewing HTML doesn't require downloading all of the document. Further, users should have the possibility to choose from online, CD, or download version of the book. The producer's challenge of preparing all of the above versions or their combinations from single document source might be further complicated by some additional requirements on "special" features accompanying the e-book such as audio, video, or software components.

It is hard to decide which type of electronic books is the best one, but there is clearly a need for a system which allows production of all of them.

MD*Book manages many output formats automatically. It is based on LaTEX and on free translation programs such as pdflatex and latex2html. In its simplest form, it runs under Linux and it has only command-line interface which offers many possiblities of influencing the behavior of the system. On the other hand, some people who don't care about understanding the system or prefer more colorful graphical user interface might appreciate MD*Book online at www.md-book.com. Here, your LaTEX source can be converted into an electronic book only by clicking on few buttons.

The inclusion of computing components is currently based on the XploRe Client/Server technology, but other computing environments could be included by adding new sets of LaTEX commands. Another example of MD*Book flexibility is the new MD*Booklet format whose many faces can be appreciated at `www.md-stat.com`.

It is not clear which of the current formats (such as Open eBook or Adobe eBook) will become the standard in the next decade, but MD*Book is a tool that can adapt and survive.

## Environmental Statistics

(Contributed Session)

Session Chair: Richard Cutler

*Friday, 8:00 am, Market West Room:*

### A Spatial Model for Chronic Wasting Disease in Rocky Mountain Mule Deer

Christopher H. Mehl, (University of Colorado at Denver), `cmehl@math.cudenver.edu`, and
Craig J. Johns, (University of Colorado at Denver), `cjohns@math.cudenver.edu`

Chronic wasting disease (CWD) causes damage to portions of the brain and nervous systems in deer and elk. The disease has been spreading rapidly throughout the Rocky Mountain Region and its economic and biological impacts have made this problem both scientifically and socially important. Previous efforts for modeling the spread of the disease have focused on simulating stochastic individual interactions between deer using standard epidemic models. We propose a hierarchical Bayesian model that captures the spatial and temporal components of the disease spread and incorporates multiple data types. Critical to our model are differential equations used to represent disease dynamics, the hierarchy which aggregates the individual interactions in both space and time, and the Bayesian formulation which naturally incorporates available data to estimate parameters in the model. The Bayesian formulation makes prediction into the future possible, which is a useful addition to disease management efforts.

*Friday, 8:20 am, Market West Room:*

### Spatial Statistics in the Presence of Location Error

John Kornak, (The Ohio State University), `jk@stat.ohio-state.edu`, and
Noel Cressie, (The Ohio State University), `ncressie@stat.ohio-state.edu`

Techniques for the analysis of spatial data have, to date, tended to ignore any effect caused by error in specifying the spatial locations at which measurements are recorded. Methodology for adjusting spatial inference in the presence of data-location error is presented for data that have a continuous spatial index (geostatistical data). New kriging equations are developed and evaluated based on a simulation study. They are also applied to remote-sensing data from the Total Ozone Mapping Spectrometer instrument on the Nimbus-7 satellite, where the location error is caused by assignment of the data to their nearest grid-cell centers. The remote-sensing data measure Total Column Ozone (TCO), which is important for protecting the earth's surface from ultraviolet and other radiation.

*Friday, 8:40 am, Market West Room:*

### Separating Signal from Noise in Global Warming

Bert W. Rust, (National Institute of Standards and Technology), bwr@cam.nist.gov

Many people still refuse to acknowledge the reality of global warming. One argument often used against it is that the global temperature record is too noisy to allow a clear determination of the signal. This paper presents two models for the signal which demonstrate that: (1) the warming in the last 146 years has been more than 8.5 times greater than the noise level, (2) the warming is accelerating, and (3) the warming is related to the growth in fossil fuel emissions. One model uses a constant rate for the acceleration and the other an exponential whose rate constant is exactly one half that of the growth in fossil fuel production. The two models can be viewed as best case and worst case scenarios for extrapolations into the future, but the data measured so far cannot reliably distinguish between them.

*Friday, 9:00 am, Market West Room:*

### Predictive Mapping of Forest Characteristics for Fire Risk Assessment

Gretchen Moisen, (US Forest Service), gmoisen@fs.fed.us,
Tracey Frescino, (US Forest Service), tfrescino@fs.fed.us,
Cheng Huang, (SAIC, US Geological Survey), huang@usgs.gov,
Jim Vogelmann, (SAIC, US Geological Survey), vogel@usgs.gov, and
Zhiliang Zhu, (US Geological Survey), zhu@usgs.gov

Maps of forest cover type and canopy height are needed for LANDFIRE, a multi-scale fire risk assessment project designed to generate intermediate-resolution data of vegetation and fire fuel characteristics for the U.S. Here we describe an evaluation study in the central Rockies of Utah, comparing tree-based methods, multivariate adaptive regression splines (MARS), and a hybrid method for mapping forest cover and canopy height on the basis of more than 2,000 forest inventory ground plots in the seven million ha mapping zone. The two forest attributes were modeled as functions of a variety of predictor variables, including: Landsat 7 Enhanced Thematic Mapper Plus (ETM+) images acquired at three different seasons; Tasseled-cap brightness, greenness, and wetness; a forest type group map; and topographic variables derived from Digital Elevation Models (DEMs); and other ancillary variables. The hybrid modeling approach showed a marked increase in overall and within forest cover type accuracies, outperforming the tree-based and MARS approaches. Little difference was seen in global performance measures of forest canopy height models, but patterns in residual plots resulting from different modeling approaches raise questions about utility of height predictions in different applications.

*Friday, 9:20 am, Market West Room:*

### Nonparametric Modeling of Soil Characteristics for Crop Models

Stephan R. Sain, (University of Colorado at Denver), ssain@math.cudenver.edu, and
Doug Nychka, (NCAR), nychka@cgd.ucar.edu

Deterministic models that predict yields for various agricultural crops require a number of inputs. One class of inputs is the water holding characteristics of the soil. These characteristics are not extensively measured, although other soil characteristics such as the composition (percentage of sand, clay, and silt) are widely known. We develop a additive model based on a thin-plate spline estimate to the transformed composition to estimate water holding characteristics. The impact of such estimates on predicted yields will be considered. Further, impacts of the hetereogenity in the soil at a specific location as well as across a region will also be examined.

## Social Networks and Statistics

(Invited Session)

Organizer and Session Chair: Jeff Solka

*Friday, 10:30 am, Weights & Measures Room:*

### Random-Effects Models for Network Dependence

Peter Hoff, (University of Washington), hoff@stat.washington.edu

One impediment to the statistical analysis of network data has been the difficulty in modeling the dependence among the observations. In the very simple case of binary network data, some researchers have parameterized network dependence in terms of exponential family representations. Accurate parameter estimation in this setting can be difficult, and the most commonly used models often display a significant lack of fit. Additionally, such models are generally limited to binary data. In contrast, random-effects models have been a widely successful tool in capturing statistical dependence for a variety of data types, and allow for prediction, imputation, and hypothesis testing within a general regression context. However, their application to network data has been limited.

We propose the use of novel random-effects structures for the statistical modeling of dependent network data. Such an approach typically proceeds by fitting a standard regression model, except that the error term is decomposed into a set of simple random effects which induce statistical dependence. We decompose the error term into sender-specific, receiver-specific, and dyad-specific random effects. The sender-specific effects, for example, can capture the positive dependence among observations having a common sender, and similarly for the receiver-specific effects. More difficult is the modeling of the dyad-specific effects, which ideally are able to capture more complicated forms of network dependence such as reciprocity, transitivity, and balance. We take a "latent similarity" approach to modeling dyadic dependence, in which the dyad-specific random effect is a simple function measuring the similarity of additional node-specific random effects. Such an approach provides a flexible strategy for the statistical modeling of network data using well known statistical tools such as regression and generalized linear models. The method also provides a model-based graphical representation of network data structure.

*Friday, 11:05 am, Weights & Measures Room:*

### Ultra-Robust and Scalable Networks Based on Hierarchies

Peter Dodds, (ISERP, Columbia University), peter.dodds@columbia.edu,
Duncan Watts, (Sociology, Columbia University), djw24@columbia.edu, and
Charles Sabel, (Law School, Columbia University), cfs11@columbia.edu

We report on the properties of a theoretical class of communication networks. We construct networks by adding links to an initial hierarchical network based on a two-parameter probability distribution. The networks produced range in form through random, team-based (links are added locally), core-periphery (links are added from the top down), to an intermediate class we term as multiscale networks. We observe certain multiscale networks to be ultra-robust in that they are both resilient to high loads of communication and remain highly connected in the unlikely event of a substantial, targeted loss of nodes. We find these networks require a minimal number of links to be added to achieve their characteristics, on the order of the number of nodes. We also show that ultra-robust networks perform well under a variety of communciation patterns and that their performance scales well with system size. Applications lie in the understanding and potentially the design of robust organizational networks.

*Friday, 11:40 am, Weights & Measures Room:*
**Statistical Models, Degeneracy and Inference for Social Networks**

Mark S. Handcock, (University of Washington), handcock@stat.washington.edu

The process of formulation and information encapsulated within social networks result in a form of "relational data". Relational data arise in many social science fields and graph models are a natural approach to representing the structure of these relations. We consider statistical and stochastic models for such graphs that can be used to represent the structural characteristics of the networks. In our applications, the nodes usually represent people, and the edges represent a specified relationship between the people. A commonly used model formulation was introduced by Frank and Strauss (1986) and derived from developments in spatial statistics (Besag 1974). These models allow for the potentially complex dependencies within relational data structures. To date, the use of graph models for networks has been limited by three interrelated factors: the complexity of realistic models, paucity of empirically relevant simulation studies, and a poor understanding of the properties of inferential methods. In this talk we discuss solutions to these limitations. We emphasize the important of likelihood-based inferential procedures and role of Markov Chain Monte Carlo (MCMC) algorithms for simulation and inference. A primary ongoing issue is the identification of classes of realistic and parsimonious models. In this regard show the unsuitability of some commonly promoted Markov models classes because they can result in degenerate probability distributions. We also consider the suitability and inference for classes of "power law" models that have been proposed for certain random graphs. The ideas are motivated and illustrated by the study of sexual relations networks with the objective of understanding the social determinants of HIV spread.

## Best of the International Association of Statistical Computing

(Invited Session)

Organizer: John Hinde
Session Chair: Wilfried Grossmann

*Friday, 10:30 am, Smokehouse/Rap of the Gavel Room:*
**Incremental Algorithms for Missing Data Imputation based on Recursive Partitioning**

Claudio Conversano, (University of Cassino), conversa@unina.it

In the framework of missing data imputation, we consider a non-parametric approach based on Information Retrieval. In particular, an incremental procedure based on the iterative use of recursive partitioning methods and a suitable Incremental Imputation Algorithm is proposed. The key idea is to define a lexicographic ordering of cases and variables so that conditional mean imputation via binary trees can be performed incrementally. A simulation study and real world applications are shown to describe the advantages and the good performance with respect to standard approaches for non-linear structures. Some possible extensions of the proposed approach to the problem of data validation will be also discussed.

*Friday, 11:05 am, Smokehouse/Rap of the Gavel Room:*
**Many Faces of a Tree**

Simon Urbanek, (University of Augsburg), simon.urbanek@math.uni-augsburg.de

Recursive partitioning trees offer a valuable tool to analyze structure in datasets. Due to their complexity visualiation methods are necessary to analyze and interpret tree models. There are many ways to display various structures contained in a tree. This paper describes different means of visualization of tree models with our prototype software, KLIMT (Klassification - Interactive Methods for Trees), for interactive graphical analysis of trees.

*Friday, 11:40 am, Smokehouse/Rap of the Gavel Room:*
**WiSP: An R Library for Simulating Wildlife Surveys**

Walter Zucchini, (University of Goettingen), wzucchi@uni-goettingen.de,
David Borchers, (University of St. Andrews), dlb@mcs.st-and.ac.uk,
Stefan Kirchfeld, (University of Goettingen), stefan.kirchfeld@sk-pism.de, and
Martin Erdelmeier, (University of Goettingen), erdelmeier@gmx.net

WiSP, Wildlife Simulation Package, is a new library of functions written in R (Ihaka and Gentleman, 1996) designed as a tool for introducing students and researchers to methods used in wildlife abundance estimation, and for experimenting with the methods using simulation. The methododology and underlying theory are described in Borchers, Buckland and Zucchini (2002).

WiSP is an object-oriented package that enables users
- to generate and view animal populations with realistically complex spatial and individual characteristics,
- to generate a variety of survey designs (such as quadrat, removal, mark-recapture and line or point transect sampling),
- to survey populations under different assumptions regarding the visibility/catchability of the individual animals,
- to compute point and interval estimates of abundance for each model.

We outline the structure of WiSP and the contents and properties of its main objects emphasising the modularity of the architecture. We demonstrate the capabilities of WiSP and illustrate how it can be used to assess the performance of the estimators, including their sensitivity to specific violations in model assumptions (such as heterogeneity), by means of simulation and graphical displays.

## Digital Government Research in Support of Federal Statistics

(Invited Session)

Organizers: Cathryn Dippo and Steve Cohen
Session Chair: Steve Cohen

*Friday, 10:30 am, Granary Room:*
**Using an Ontology as Generalized Metadata Schema for Access to Distributed Heterogeneous Data Sources**

Edward Hovy, (Information Science Institute University of Southern California), hovy@isi.edu

This presentation describes the Energy Data Collection (EDC) project. We merge a large general-purpose ontology and a more focused domain model and embedded the result into a system for supporting user access to over 50,000 tables of information about gasoline price and production, obtained from the Energy Information Administration, the Bureau of Labor Statistics, the Census Bureau and the California Energy Commission. The source data was provided in a variety of formats, including Microsoft Access spreadsheets, pdf and html pages, and raw text files. An inportant focus of the work was using the merged ontology/domain model as a generalized metadata schema.

*Friday, 11:05 am, Granary Room:*
**Interfaces to a Statistical Knowledge Network**

Gary Marchionini, (University of North Carolina at Chapel Hill), march@ils.unc.edu

As a result of national and local government efforts to leverage WWW technology, the public has increasingly better access to government statistical data. Two significant challenges to promoting appropriate usage of this data are to make the data understandable by people with diverse interests and statistical expertise, and to integrate relevant data from different sources and agencies. As a first step toward a statistical knowledge network that supports access and use, we focus on creating and testing human-computer interfaces that serve to integrate different data sources and provide on-demand help. In this talk, examples of highly interactive interfaces to find US federal agency data and understand statistical concepts relevant to that data will be discussed. These interfaces depend on well-structured underlying metadata and our approaches to integrating cross-agency data underneath the interface will also be discussed.

*Friday, 11:40 am, Granary Room:*
**New Approaches to Mobile Computing for Field Data Collection**

Sarah Nusser, (Iowa State University), nusser@iastate.edu

Recent research in mobile computing for survey data collection has focused on the use of geospatial data formats in small screen environments, including augmented vision and wearable computers, and on flexible infrastructure models that take into account features associated with a data collection setting. We will discuss studies that explore how map data and GPS position indicators presented in mobile computing environments affect the ability of field data collection staff to plan work and locate sample units. Results indicate that multiple tools are needed to address the variability that exists in spatial abilities across field staff. This can be accomplished by presenting a variety planning and navigation aids on the map display. This research is being used to propose infrastructure models for delivering usable geospatial data products and functions to the field via adaptable systems that accommodate a variety of field computing and application settings.

## Smoothing and Nonparametric Feature Detection

(Invited Session)

Organizer and Session Chair: Michael Minnotte

*Friday, 10:30 am, Market Street Room:*
**A SiZer Analysis of IP Flow Start Times**

J. S. Marron, (UNC), marron@email.unc.edu,
Felix Hernandez–Campos, (UNC), fhernand@cs.unc.edu, and
F. D. Smith, (UNC), smithfd@cs.unc.edu

The SiZer technique is used to study the homogeneity of a point process of Internet traffic flow start times. It is seen that a homogenous Poisson process is an inappropriate model, because it does not yield observed statistically significant burstiness. Some Weibull waiting processes gives better, but still inadequate performance. A clustered Poisson process gives the best fit.

*Friday, 11:05 am, Market Street Room:*
**Longitudinal Kernel Regression**

Naisyin Wang, (Texas A&M University), nwang@stat.tamu.edu,
Raymond J. Carroll, (Texas A&M University), carroll@stat.tamu.edu,
Xihong Lin, (U. Of Michigan), xlin@sph.umich.edu, and
Ziding Fend, (Fred Hutchinson Cancer Center), zfeng@fhcrc.org

There has been a substantial recent interest in investigating the performance of kernel regression estimator for longitudinal data. Most approaches adopt the strategy of ignoring the within-subject correlation structure. When the cluster sizes remain fixed, a result supporting the use of this "working independence" strategy indicates that under the conventional estimation procedure, a correct specification of the correlation structure actually diminishes the asymptotic efficiency. In this presentation, I will discuss an alternative kernel estimating equation that accounts for the within subject correlation. The major gain by the new approach is at variation reduction. For nonparametric curve estimation, the variance of the proposed method is uniformly smaller than that of the most efficient working independence approach. Under the framework of marginal generalized partially linear models, the new estimator is semiparametric efficient in the Gaussian case, and is more efficient than the working independence estimator in non–Gaussian cases.

*Friday, 11:40 am, Market Street Room:*

### Semiparametric Regression Smoothing and Feature Detection

Michael G. Schimek, (Karl-Franzens-University Graz, Austria),
michael.schimek@uni-graz.at

Let us assume a regularly spaced time series with white noise or weak stationary autoregressive errors of known order. Correct nonparametric estimation of a smooth trend in a series of dependent observations asks for a regression technique which takes care of the specific error structure. Such an approach has been given in Schimek and Schmaranz (1994). Here we extend it to a semiparametric concept in which one or more artificial dummy input series are introduced to analyze certain features in the time series which are of interest beyond long-term trend. Unbiased partial spline fitting (Schimek, 2002) is a recent approach to evaluate such a model. An open problem since Schimek and Schmaranz (1994) is the choice of the degree of smoothing, more so in the semiparametric context where dummy (feature) testing is involved. Whilst the trend is fitted nonparametrically by a smoothing spline, features are tested parametrically via a simple approximately N(0,1) or F-distributed test statistic.

Criteria are derived and estimates given to calculate the significance trace (introduced in a nonparametric setting by Azzalini and Bowman, 1993). The p-values are plotted as a function of the smoothing parameter across a typical range of values. Doing so we can study test significance as a function of the degree of nonparametric smoothing, thus having a systematic handle for feature detection. The main complication here is computational costs. Finally we illustrate the approach on a real example.

## Safety and Security

(Contributed Session)

Session Chair: Karen Kafadar

*Friday, 10:30 am, Market West Room:*

### An Economic Index for Evaluating Traffic Safety

Michael Conerly, (University of Alabama), mconerly@cba.ua.edu,
J. Michael Hardin, (University of Alabama), mhardin@cba.ua.edu,
Wade Watkins, (University of Alabama), wwatkins@cba.ua.edu
Chunyao Feng, (University of Alabama), and
Bo Hong (University of Alabama)

Data analysis methodologies that assist in highlighting patterns of traffic safety offer policy and decision makers valuable tools for accomplishing their tasks of developing and recommending interventions to solve traffic safety problems. Using standard statistical procedures, this paper will develop an economic index useful for comparing the economic impact of traffic crashes in the 67 counties in Alabama. The index adjusts for various couunty and demographic factors such as population, area of county, proportion of teenage drivers in the county, etc. The data used was extracted from the accident database created by the state department of transportation. This methodology may be applied to other states and/or other geographic regions.

*Friday, 10:50 am, Market West Room:*
### Bayesian Inductively Learned Modules for Safety Critical Systems

Jonathan E. Fieldsend, (School of Engineering and Computer Science, University of Exeter), J.E.Fieldsend@ex.ac.uk,
Trevor C. Bailey, (Mathematical Sciences, U. of Exeter), T.C.Bailey@ex.ac.uk,
Richard M. Everson, (Engineering and Computer Science, U. of Exeter), R.M.Everson@ex.ac.uk,
Wajtek J. Krzanowski, (Mathematical Sciences, U. of Exeter), W.J.Krzanowski@ex.ac.uk,
Derek Partridge, (Engineering and Computer Science, U. of Exeter), D.Partridge@ex.ac.uk,
and
Vitaly Schetinin, ( Engineering and Computer Science, U. of Exeter), V.Schetinin@ex.ac.uk

This work examines the use of Bayesian inductively learned software modules for safety critical systems. Central to the safety critical application is the desire to generate confidence measures associated with predictions. This is achieved in this study by casting the problem in a Bayesian formulation, and is implemented using reversible jump Markov Chain Monte Carlo (RJ-MCMC). We use conventional and novel classification architectures, including logistic discriminants, probabilistic k-nn and radial basis function networks. Results from these methods are illustrated on real life critical systems, including medical trauma data. We report results on the trade-off between model complexity and the width of the posterior predictive probability.

*Friday, 11:10 am, Market West Room:*
### Waypoint Analysis for Command and Control

Mark Irwin, (Ohio State University), irwin@stat.ohio-state.edu,
David Wendt, (Battelle Memorial Institute), wendtd@battelle.org, and
Noel Cressie, (Ohio State University), ncressie@stat.ohio-state.edu

Command and Control (C2) is a broad field, providing a wide variety of options for research in statistical methodology and application. A basic problem in C2 is the ability to track an enemy object in the battlespace and to forecast its future position. However, particularly in a military context, a mobile object is often attempting to reach a pre-designated location at a pre-designated time. If one is able to assume that the enemy object is headed for an unknown location, or waypoint, then the statistical problem changes fundamentally. A Bayesian hierarchical modeling approach, linking the movement model with the waypoint, will be taken. For comparison, an ad hoc least squares estimator, based on triangulation, is also developed. Properties of the estimators and computational considerations will be discussed in terms of a small simulation study.

*Friday, 11:30 am, Market West Room:*
### Continually Improving Stream Analysis for Network Security

Nancy J. McMillan, (Battelle), mcmillann@battelle.org,
Douglas D. Mooney, (Battelle), mooneyd@battelle.org, and
Dave A. Burgoon, (Battelle), burgoon@battelle.org

In many real-world environments, events happen at irregular intervals and measurements describing those events are recorded, e.g., network connection attempts. The flow of measurements thus generated is stream data. The pace of real world events, which is not controlled, governs the rate at which stream data flows. The real-time management and use of stream data for decision-making and/or characterization is complicated by the variable flow rate inherent in this data. By nature, these activities require data management and data processing based on algorithms. Data management and processing takes time; the amount of time is governed by the complexity of the algorithms employed. Typically more complex and time-consuming algorithms are only considered when they provide superior decision-making ability or superior characterization. However, if new data cannot be handled as quickly as it is generated, real-time management and use is not occurring. Thus, there is a natural trade-off between algorithms that store and process data quickly enough to keep up with the flow of stream data and algorithms that provide a sufficiently accurate decision or characterization. Continually improving stream analysis (CISA) is proposed as a mechanism for managing the trade-off between providing sufficiently accurate decisions/characterizations and keeping up with the flow of stream data. The real-time stream data monitoring features that are provided by CISA are: 1. The algorithm always provides a current decision or characterization. 2. The precision/accuracy of the current decision or characterization improves when there is more processing time available relative to the rate of data flow, i.e., more processors, faster processors, and slower data flow all translate to more precise/accurate decisions or characterizations. 3. The algorithm scales automatically to optimize accuracy/precision of the current decision or characterization as a function of data flow rate. In this work, concepts for CISA are realized in the framework of a cyber security example. Specifically, a CISA intruder detection system (IDS) is developed, which monitors firewall data. The IDS developed is a dynamic classification/characterization tool that first identifies groups of sources by common behavior patterns then characterizes the behavior of the groups identified over time. As new intruder behaviors emerge, they are captured by the appearance of new groups or the migration in behavior patterns of existing groups.

*Friday, 11:50 am, Market West Room:*
**A Micro-scale Epidemiological Simulation for Management of Disease Outbreaks**

Sid Baccam, (Los Alamos National Laboratory), pbaccam@lanl.gov,
Stephen Eubank, (Los Alamos National Laboratory), eubank@lanl.gov, and
Catherine Macken, (Los Alamos National Laboratory), cmacken@lanl.gov

We have created an agent-based simulation of 1.6 million synthetic individuals to study medical interventions of infectious disease outbreaks. The simulation involves population mobility of people from their homes to work, shopping, etc., a transmission model of the pathogen (in this case, influenza virus), and the viral/immune kinetics within infected individuals. Case studies will be presented that test different vaccination strategies and antiviral therapy options. Real-world household studies of influenza virus infection have been analyzed, and their implications on our simulation will be discussed.

## Public Health Preparedness and Response in Crisis

(Invited Session)

Organizer and Session Chair: Jimmie D. Givens

*Friday, 1:45 pm, Weights & Measures Room:*
**Using Design-Based Adaptive Sampling Procedures in Site Decontamination**

Myron J. Katzoff, (National Center for Health Statistics), mjk5@cdc.gov,
Abera Wouhib, (National Center for Health Statistics), aqw3@cdc.gov, and
Joe Fred Gonzalez, Jr., (National Center for Health Statistics), jfg2@cdc.gov

In this paper, we consider the application of finite-population design-based sampling procedures in a spatial context to decontamination of a site where there is a significant public health risk of anthrax exposure. Through computer simulation, we study the properties of adaptive sampling procedures employed in the search of a bounded three-dimensional space that serves as a model of the site. For a finite set of designs, we compare the operational efficiency of procedures, as measured by percent of contamination eliminated, and examine the variation in detection probabilities with choices of selection-unit parameters, cloud-density and design complexity.

*Friday, 2:30 pm, Weights & Measures Room:*
**Game Theory and Risk Analysis for the Smallpox Threat**

David Banks, (U.S. Food and Drug Admin.), banksd@cber.fda.gov

Federal agencies must allocate their limited resources for counterterrorism programs in the most effective ways possible. These decisions entail both statistical risk analysis, to account for uncertainties in the costs and benefits of particular efforts, and game theory, to account for the fact that terrorists adapt their attacks in response to homeland defense initiatives. This talk describes a procedure that uses risk analysis to generate random payoff matrices for game theory solution, and then pools the solutions from multiple realizations to estimate the probability that a given play is optimal. The strategy is illustrated for risk management of the threat of smallpox attacks.

## Best of the Journal of Computational and Graphical Statistics

(Invited Session)

Organizer and Session Chair: David Scott

*Friday, 1:45 pm, Smokehouse/Rap of the Gavel Room:*
**Penalized Survival Models and Frailty**

V. Shane Pankratz, (Mayo Clinic - Division of Biostatistics), pankratz.vernon@mayo.edu,
Patricia M. Grambsch, (University of Minnesota School of Public Health - Division of Biostatistics), pat@biostat.umn.edu, and
Terry M. Therneau, (Mayo Clinic - Division of Biostatistics), therneau@mayo.edu

Interest in the use of random effects in survival analysis settings has been increasing. However, the computational complexity of such frailty models has limited their general use. While fitting frailty models has traditionally been difficult, parameter estimation in penalized Cox semi-parametric and parametric regression models can be done using simple extensions of the standard algorithms for fitting non-penalized models. We demonstrate that solutions for gamma shared frailty models can be obtained exactly via penalized estimation. Gaussian frailty models are also closely linked to penalized models. Therefore, fitting frailty models with penalized likelihoods can be made quite efficient by taking advantage of computational methods available for penalized models. We have implemented penalized regression for the coxph function of S-plus. In this presentation, we outline the links between frailty models and penalized likelihood methods, and illustrate use of the algorithms implemented within the coxph S-Plus function with several examples. Of particular interest, we have successfully used these methods on data from a large study of the genetic epidemiology of breast cancer.

*Friday, 2:20 pm, Smokehouse/Rap of the Gavel Room:*
**Adaptive Order Selection for Spline Smoothing**

Randy Eubank, (Texas A&M University), eubank@stat.tamu.edu,
Chunfeng Huang, (North Dakota State University), chunfeng.huang@ndsu.nodak.edu, and
Suojin Wang, (Texas A&M University), sjwang@stat.tamu.edu

Computational methods are presented for spline smoothing that make it practical to compute smoothing splines of degrees other than just the standard cubic case. Specifically, an order n algorithm is developed that has conceptual and practical advantages relative to classical methods. From a conceptual standpoint, the algorithm uses only standard programming techniques that do not require specialized knowledge about spline functions, methods for solving sparse equation systems or Kalman filtering. This allows for the practical development of methods for adaptive selection of both the level of smoothing and degree of the estimator. Simulation experiments are presented that show adaptive degree selection can improve estimator efficiency over the use of cubic smoothing splines. Run time comparisons are also conducted between the proposed algorithm and a classical, band-limited, computing method for the cubic case.

*Friday, 2:55 pm, Smokehouse/Rap of the Gavel Room:*
**An Adaptive Spatial Scan Density Estimation Method**

Ramani S. Pilla, (Department of Statistics, Case Western Reserve University), pilla@po.cwru.edu,
Peng Tao, (Accu Image Diagnostics Corporation), peng_tao2000@yahoo.com, and
Carey Priebe, (Department of Mathematical Sciences, Johns Hopkins University), cep@jhu.edu

Spatial scan density (SSD) estimation via mixture models is an important problem in the field of spatial statistical analysis and has wide applications in image analysis. The "borrowed strength" density estimation (BSDE) method via mixture models (Priebe, 1996) enables one to estimate the local probability density function in a random field wherein potential similarities between the density functions for the subregions are exploited. This article proposes an efficient method for SSD estimation by integrating the borrowed strength technique into the alternative EM framework (Pilla & Lindsay, 2001) which combines the statistical basis of the BSDE approach with the stability and improved convergence rate of the alternative EM methodology. In addition, we propose an adaptive SSD estimation method that extends the aforementioned approach by eliminating the need to find the posterior probability of membership of the component densities afresh in each subregion. Simulation results and an application to the detection and identification of man-made regions of interest in an unmanned aerial vehicle imagery experiment show that the adaptive method significantly outperforms the BSDE method. Other applications include automatic target recognition, mammographic image analysis and minefield detection.

## Infrastructure Security

(Invited Session)
Organizers: Dale Anderson and Sandra Thompson
Session Chair: Sandra Thompson

*Friday, 1:45 pm, Granary Room:*
**Energy Infrastructure Vulnerability Assessments**

Jeff Dagle, (PNNL), jeff.dagle@pnl.gov

The Department of Energy (DOE) vulnerability and risk assessment program was established to assist energy infrastructure providers in assuring the continued delivery of their critical services. These vulnerability assessments differ from those that are provided by other organizations in process, focus, scope and access to expertise. An assessment approach leveraged from other assessment methodologies has been developed that is unique. The focus is vulnerabilities of critical process control systems in the context of a broader enterprise security assessment. Additionally, both physical and cyber security are evaluated, along with an analysis of threat, impact, and overall risk characterization to provide a framework for prioritizing recommendations. The DOE objective is to enable the energy infrastructure provider to enhance its security posture, with the understanding that these organizations are stewards of infrastructures with significant national importance.

*Friday, 2:20 pm, Granary Room:*
**PNNL & International Border Security**

William C. Cliff, (PNNL), william.cliff@pnl.gov

PNNL's International Border Security Training program (INTERDICT/RADACAD) trains and equips international border-enforcement officials to detect, identify, interdict and investigate all aspects of smuggling related to the proliferation and diversion of materials, commodities and components associated with the development and deployment of nuclear, chemical and biological weapons and their related missile delivery systems. This presentation will attempt to point out how specialized training and todays advanced detection technologies are assisting in the worlds Commerce Infrastructure Security. The presentation will also attempt to show the need for statistical targeting of commerce to more effectively use the limited current search resources.

*Friday, 2:55 pm, Granary Room:*
**Electricity Infrastructure Security**

Thomas Kropp, (EPRI), tkropp@epri.com

Todays power system relies on an increasingly stressed infrastructure. This is due to several factors. Infrastructure expansion has not kept up with demand: generation & transmission capacity margins are shrinking. Transition to competition is creating new demands. Technology can meet these demands, but uncertainties on ROI are discouraging investments. Many distribution systems have not been updated with current technology. There has been a proliferation of distributed resources, yet few have been connected to the grid. The national infrastructure security assessment adds to concern.

The threat of deliberate attacks on the power system adds a new dimension to concerns about the infrastructure. Transformers, line reactors, series capacitors, and transmission lines are all vulnerable to attack and it is impractical to protect all of our diverse and dispersed assets. There are over 200,000 miles of EHV lines and over 6,500 Transformers in the Eastern Interconnect alone. The power system is intrinsically connected to other infrastructures, such as natural gas pipelines and compressor stations, hydroelectric dams, rail lines, and telecommunications. In addition, our power system is a North American system, serving Canada as well as the United States.

Given our reliance on electricity to drive all of the devices upon which we are now dependent, we must undertake significant efforts to improve the security, resilience, and dependability of our power system. This talk will discuss technology required to move the North American power system toward these goals.

## Refereed Contributed Papers I: Statistical Computing

(Refereed Contributed Session)

Session Chair: Tim Hesterberg

*Friday, 1:45 pm, Market Street Room:*
**RGL: A R-library for 3D Visualization with OpenGL**

Oleg Nenadic, (University of Goettingen), onenadi@uni-goettingen.de,
Daniel Adler, (University of Goettingen), dadler@gwdg.de, and
Walter Zucchini, (University of Goettingen), wzucchi@uni-goettingen.de

RGL is a library of functions that offers three-dimensional, real-time visualization functionality to the package R (Ihaka and Gentleman, 1996), thereby ameliorating a shortcoming in the current version of R as well as most other statistical software, namely its inability to allow the user to conveniently generate interactive 3D displays.

Graphical visualization techniques, especially two-dimensional displays such as scatterplots and histograms, are routinely used in statistical data analysis to explore datasets in order to reveal their properties and structure. Since 3D objects need to be projected on a 2D display, special navigation facilities are required for gaining an insight into 3D relationships. Features such as lighting, alpha-blending, texture-mapping and fog-effects are useful for enhancing the illusion of three-dimensionality. Additional desirable features for interactive data analysis in 3D include the ability to rotate objects and to zoom in/out so as to examine details of an object, or alternatively, to view it from a distance.

The goal of the project described here was to provide a dll interface (written in C++) from R to OpenGL which then acts as a '3D engine'. This way, high-level plotting functions can be written in R, which use primitives (points, lines, triangles, spheres, surfaces etc. in 3D space) provided by the library. RGL is developed with long-term goals in mind, resulting in e.g. cross-platform portability and reduced complexity due to modularization and object-oriented design. Furthermore, the syntax of the RGL commands has been based on that of the related and familiar standard R commands, thus ensuring that users familiar with the latter can quickly learn the usage of RGL. This paper describes the structure of the RGL library and illustrates its capabilities by means of a number of examples.

*Friday, 2:10 pm, Market Street Room:*
**Maximum Entropy Constructive Ensembles for Time Series Analysis**

H. D. Vinod, (Fordham University), vinod@fordham.edu

The ensemble plays a key role as the notional population for the observed time series. I propose a new method of constructing the ensemble by using maximum entropy (ME) methods. The ME distribution satisfies the mean preserving constraint by construction and is computer intensive. My seven-step algorithm for constructing ensembles is designed to satisfy the ergodic theorem and Doobs theorem, without assuming stationarity and without using asymptotics. Proposed methods are particularly convenient for short nonstationary time series and can potentially simplify several inference problems in time series analysis. Three examples illustrate them. A consumption function example explicitly shows that: (i) the constructed ensemble retains the basic shape and dependence structure of autocorrelation function (acf) and partial autocorrelation function (pacf) of the original time series, (ii) one can avoid shape-destroying transformations (differencing) and the underlying need for achieving stationarity, and (iii) one can provide confidence intervals for coefficients of lagged dependent variables. A demand function example shows that traditional inference methods are consistent and conservative when viewed in terms of the proposed ensemble.

*Friday, 2:35 pm, Market Street Room:*
**The Quickest Sequential Detection of Intrusions in Computer Networks**

Boris Rozovskii, (Department of Mathematics, University of Southern California),
rozovski@math.usc.edu,
Rudolf Blazek, (Department of Mathematics, USC), blazek@math.usc.edu,
Hongjoong Kim, (Department of Mathematics, University of North Carolina at Charlotte),
hjkim@uncc.edu, and
Alexander Tartakovsky, (Department of Mathematics, USC), tartakov@math.usc.edu

Rapid response, minimal false alarm rate, and the capability to detect a wide spectrum of attacks are the crucial features of intrusion detection systems. Proper choice of the observed network flow and resource usage characteristics is an essential aspect of the development of intrusion detection methods. Once the set of observables is decided upon, sequential change-point detection algorithms can be used to optimize the detection delay for a desired given false alarm rate. In this paper, based on the advanced change-point detection methods, we propose an efficient anomaly detection system that detects denial-of-service attacks at extremely high speeds (up to SONET OC-192) with minimal detection delay for a given low false alarm rate. These methods employ statistical analysis of data from multiple layers of the network protocol for detection of subtle traffic changes. The sequential detection algorithm is nonparametric and utilizes thresholding of a test statistic to achieve a fixed rate of false positives. The proposed constant false alarm rate detector has the following important features: it is self-learning and adapts to various network loads and usage patterns; it allows for the detection of attacks with small average delay while controlling false alarm rate at a pre-specified level; and it is computationally simple. The results of the theoretical and experimental studies with the use of a network simulator testbed are presented.

*Friday, 3:00 pm, Market Street Room:*
**Implementing Legacy Statistical Algorithms in a Spreadsheet Environment**

Stephen W. Liddle, (Brigham Young University), liddle@byu.edu, and
John S. Lawson, (Brigham Young University), lawson@byu.edu

Over the years creative researchers in the field of statistics have published many useful and freely available FORTRAN programs implementing novel statistical computations. But it is becoming less common to find a FORTRAN compiler installed on the average workstation, making this prior work less accessible to the masses who rely on personal computers running Windows. We propose a methodology for converting legacy FORTRAN algorithms to run as VBA macros in the Excel environment. Using our approach, moderately literate programmers can effectively migrate legacy algorithms to the ubiquitous Excel spreadsheet tool, where users will again have ready access to past algorithm contributions.

## Bioinformatics

(Contributed Session)

Session Chair: Bart Weimer

*Friday, 1:45 pm, Market West Room:*
**A Bayesian Mixture Model for Bayesian Gene Expression**

Kim–Anh Do, (U.T. M.D. Anderson Cancer Center), kim@mdanderson.org,
Peter Mueller, (U.T.M.D. Anderson Cancer Center), pm@mdanderson.org, and
Feng Tang, (U.T. M.D.Anderson Cancer Center), ftang@mdanderson.org

We propose model-based inference for differential gene expression, using a non-prametric Bayesian probability model for the distribution of gene intensities under different conditions. The probability model is essentially a mixture of normals. The resulting inference is similar to the empirical Bayes approach proposed by Efron et al. (JASA, 2001). The use of fully model-based inference mitigates some of the necessary limitations of the empirical Bayes method. However, the increased generality of our method comes at a price. Computation is not as straightforward as in the empirical Bayes scheme. But we argue that inference is no more difficult than posterior simulation in traditional nonparametric mixture of normal models. We illustrate the proposed method in two examples, including a simulation study and a microarray experiment to screen for genes with differential expression in colon cancer versus normal tissue.

*Friday, 2:05 pm, Market West Room:*
**A Simple Approach to Accomodating Interactive and Batch Processes on a Bioinformatics Cluster**

Warren M. Snelling, (USDA-ARS-USMARC), snelling@email.marc.usda.gov,
John W. Keele, (USDA-ARS-USMARC), keele@email.marc.usda.gov, and
Gregory P. Harhay, (USDA-ARS-USMARC), harhay@email.marc.usda.gov

Compute clusters, consisting of several tightly networked nodes assembled from commodity PC hardware, offer substantial computing power for relatively little expense. Efficiently utilizing this power for biological computations can be challenging if the cluster serves several scientists having different needs. Bioinformaticists, managing and analyzing large data sets, may split large jobs into many smaller tasks distributed across the cluster so each node solves pieces of the problem. Biologists concentrating on a few sequences will see little advantage from having computations parallelized across several nodes; they need instant access without waiting for a extended batch jobs to finish.

Providing high performance parallelization for large jobs and high availability access to interactive users can be accomplished with a relatively simple strategy combining batch programming and a prioritized queueing system. Our implementation utilizes open source Grid Engine software for the queueing system and a perl interface to the Grid Engine that we developed to reduce programming effort. On a Linux cluster, we simultaneously provide an interactive web-based BLAST server for single-sequence searches and parallel processing capabilities for linkage mapping, batch BLAST, and other analyses of large data sets.

*Friday, 2:25 pm, Market West Room:*
**Selecting an Optimal Rejection Region for Multiple Testing: A Decision-Theoretic Alternative to FDR Control, with an Application to Microarrays**

David R. Bickel, (Medical College of Georgia), dbickel@mail.mcg.edu

As a measure of error in testing multiple hypotheses, the decisive false discovery rate (dFDR), the ratio of the expected number of false discoveries to the expected total number of discoveries, has advantages over the false discovery rate (FDR) and positive FDR (pFDR). The dFDR can be optimized and often controlled using decision theory, and some previous estimators of the FDR can estimate the dFDR without assuming weak dependence or the randomness of hypothesis truth values. While it is suitable in frequentist analyses, the dFDR is also exactly equal to a posterior probability under random truth values, even without independence. The test statistic space in which null hypotheses are rejected, called the rejection region, can be determined by a number of multiple testing procedures, including those controlling the family-wise error rate (FWER) or the FDR at a specified level. An alternate approach, which makes use of the dFDR, is to reject null hypotheses to maximize a desirability function, given the cost of each false discovery and the benefit of each true discovery. The focus of this method is on discoveries, unlike related approaches based on false nondiscoveries. A method is provided for achieving the highest possible desirability under general conditions, without relying on density estimation. A Bayesian treatment of differing costs and benefits associated with different tests is related to the weighted FDR when there is a common rejection region for all tests. An application to DNA microarrays of patients with different types of leukemia illustrates the proposed use of the dFDR in decision theory. Comparisons between more than two groups of patients do not call for changes in the results, as when the FWER is strictly controlled by adjusting p-values for multiplicity.

Full text: www.arxiv.org, www.mathpreprints.com, www.davidbickel.com

*Friday, 2:45 pm, Market West Room:*
**Statistical Methods for Spot Detection with Macroarray Data**

Adele Cutler, (Utah State University), adele@math.usu.edu,
Andrejus Parfionovas, (Utah State University), andrej@cc.usu.edu,
Bart Weimer, (Utah State University), milkbugs@cc.usu.edu, and
Yi Xie, (Johns Hopkins University), yixie@cc.usu.edu

We describe a statistical method for detecting macroarray spots and extracting intensity information. The method can be used to approximate the true intensity in the presence of truncation of the data due to limitations in the dynamic range of the photographic film. The method is illustrated on experimental data, including some dilution experiments.

## Statistical Issues in Computer Security

(Invited Session)

Organizer and Session Chair: David Marchette

*Friday, 4:00 pm, Weights & Measures Room:*
**Worm Propagation on Graphs with Heavy-tailed Degree Distribution**

Stephan Bohacek, (University of Delaware), bohacek@eecis.udel.edu

Email worms spread over graphs defined by email address books. In order to study the propagation of email worms, it is necessary to understand the graph over which they spread.

The email address book graph and other graphs such as the World Wide Web and Internet routers have received much attention recently. There has been extensive work suggesting that for these graphs, the degree of each node is distributed according to a heavy-tailed distribution. Other work has shown that the World Wide Web and Internet router graph can be decomposed into two basic parts: a highly connected core and lightly connected tendrils. This structure provides insight into worm propagation. This paper will explore these graphs and show how heavy-tailed degree distribution typically leads to a highly connected core. However, we will also see how graphs such as the Internet router graph are not typical. For example, the core is not nearly as connected as the theory leads one to expect. Furthermore, the growth function (the number of nodes visited as a function of the number of "hops" from starting node) is significantly different from that of the typical graph with heavy-tailed degree distribution. Notwithstanding the Internet router graph's peculiarity, to a worm, the Internet graph appears similar to the typical heavy-tailed degree distribution. A slightly different notion of connectivity leads to an explanation for this behavior.

*Friday, 4:25 pm, Weights & Measures Room:*
**User Profiling for Intrusion Detection in Windows NT**

Tom Goldring, (U.S. DOD), tgo@tycho.ncsc.mil

In User Profiling, we observe the normal behavior of computer users and from this, seek to automatically learn models that characterize this behavior. Then for a new session, these models are used to either authenticate the login name, or to identify a malicious insider. A related problem is Program Profiling, in which models for normal activity of an application program are learned, then used to identify attacks. This is a somewhat easier problem because humans do not come with "specs", so compared to programs, our behavior is infinitely less predictable. In fact, a certain level of anomalous activity in human behavior is inevitable and must be taken into account.

Most if not all published work on this subject has used command line activity as its data source, collected on a Unix system. In this environment there are multiple ways to do most things, leaving much room for individual expression, yet even so the reported results have been less than stellar. Now consider today's point and click world, where command line activity is virtually nonexistent. Even worse, the Windows suite of interlinked applications provides a "path of least resistance", with the result that people look more alike than ever. Add to this the fact that much of the activity occurring on a host, especially if it's networked, is generated by the operating system and not user related. This requires massive filtering, but how to it accurately can be far from obvious. These considerations underscore the inherent difficulty of the problem.

For nearly two years we have been monitoring real users doing their daily work on an operational Windows NT network. This talk will describe the data we collect and methods we have used to analyze it, and present results obtained to date.

*Friday, 4:50 pm, Weights & Measures Room:*
**A Stochastic Model of Computer Intrusions for Evaluation and Exercises**

Robert P. Goldman, (SIFT, LLC), rpgoldman@sift.info

Research and practice in computer intrusion defense is hampered by the great difficulty of conducting repeatable experiments and exercises. We have drawn on recent work in artificial intelligence action modeling to design a flexible simulation environment for computer intrusions. A key component of this model is a relatively intelligent simulated attacker, built using techniques of goal-directed procedure invocation. We hope that the resulting approach will provide a good trade-off between fidelity of modeling and the ability to repeatably exercise computer security techniques.

*Friday, 5:15 pm, Weights & Measures Room:*
**Multi-Level Monitoring and Fuzzy Clustering to detect Cyber Attacks**

Dipankar Dasgupta, (The University of Memphis), dasgupta@memphis.edu,
Jonatan Gomez, (The University of Memphis), jgomez@memphis.edu, and
Fabio Gonzalez, (The University of Memphis), fgonzalz@memphis.edu

The paper investigates behavior-based techniques for detecting intrusion/anomalies. Specifically, this approache monitors data at multiple levels (from packet to user-level) in order to determine correlation among the observed parameters for efficiently detecting cyber attacks. In particular, we developed an efficient clustering and recognition techniques that can characterize the abnormal behavior to determine cyber attacks.

We applied techniques based on modeling the normal behavior (positive characterization) based on a set of normal usage data. Then, we used normal usage data to build models for abnormal behavior (negative characterization) in complement space inspired by the natural immune system.

Our work attempts to handle the uncertainty inherent in the usage data and in the decision making process using fuzzy sets to describe the input parameter space, and the normal behavior patterns. Then used fuzzy rules to build a decision support system for the detection of cyber attacks.

In particular, we investigated clustering methods that allow the data to belong to several clusters with different fuzzy membership degrees, can yield an accurate model even in the presence of noise or outliers, can automatically determine the number of clusters, and can yield elastic models that can easily adapt to fluctuations in the monitored system behavior.

## Interactive GeoGraphics for the Web

(Invited Session)

Organizer and Session Chair: Robert M. Edsall

*Friday, 4:00 pm, Smokehouse/Rap of the Gavel Room:*
**Integrated Climate Database**

Dan Dansereau, (Utah State University), dad@cc.usu.edu, and
Robert R. Gillies, (Utah State University), rgillies@nr.usu.edu

Agencies (e.g., The Bureau of Land Management) currently have to access climate data from many different web sources. This presents a number of problems: apart from the search being a time consuming process, data is often incomplete, difficult to query and reference, and generally not in a format for straightforward ingestion into spreadsheets or a GIS framework. Moreover, existing websites are often not tailored for the varying needs of multiple users.

An initiative at Utah State University has been to build a "one stop" warehouse integrated climate database (ICB) that is web accessible. This database is federated in the sense that it takes advantage of existing data-feeds to collect near real-time data in addition to the existing historical information. The data is accessible via a GIS interface that allows for querying by user definable maps. Products (e.g., reports and graphs), either as pre-formatted or user designable, are available via the web xml interface. User requested data for download can be packaged in numerous formats, e.g., spreadsheet, GIS or simple CSV formatting. In summary, the ICB is web accessible, has user-friendly navigation and data output, high quality graphs, reports and graphs, ingests data from many sources and is available in one location. It is our intention to demonstrate this functionality in the session

*Friday, 4:45 pm, Smokehouse/Rap of the Gavel Room:*
**Web Cartography for Municipal Government: An Accessibility Case Study**

Robert M. Edsall, (Arizona State University), robedsall@asu.edu

The interactive 2002 Green Map of Tempe, AZ, was influenced by choices typical of all cartographic design (color, symbolization, generalization, layout, balance, etc.). In addition, the creation of the web-based map of ecologically significant sites in and around Tempe was also affected by considerations more exclusive to dynamic maps open to all users. This paper will report on several of these design requirements, including ways to maximize limited screen real estate, methods to accommodate user customization, and means of providing access to the maps to the disabled. This final consideration will be discussed in detail; a 1998 act of the U.S. Congress gave citizens and government employees the right to sue public agencies in federal court if those agencies fail to provide equal access to electronic information to those with disabilities. The Green Map is the first of a series of collaborations between the Arizona State Department of Geography with the municipal government of the City of Tempe. The City is required to address accessibility issues with the information they post on their web site, and this paper will discuss the strategies that were employed (or still need to be employed) by cartographers at Arizona State to bring the Green Map and other Internet mapping projects into compliance with the act.

## Design and Statistical Computation

(Contributed Session)

Session Chair: Bill Shannon

*Friday, 4:00 pm, Granary Room:*
**Application of Simulated Annealing to D-optimal Design for Polynomial Regression with Correlated Observations**

Zewen Zhu, (Utah State University), sl4sv@math.usu.edu, and
Daniel C. Coster, (Utah State University), coster@math.usu.edu

Simulated annealing (SA) is useful for solving, at least approximately, complex optimization problems for which an analytical solution is not known. We use SA to find D-optimal design(s) for polynomial regressions with correlated observations and multiple correlation structures in the exact design setting. For given state (design) spaces, cooling schedules, and perturbation schemes, the SA results show that the D-optimal design(s) for polynomial regression with correlated observations (i) depend on the correlation structures, and (ii) these D-optimal designs are not, in general, the same as the D-optimal designs with uncorrelated observations. In particular, both the number and location of the optimal regression support points differ between the correlated and uncorrelated cases, and the order of observations within blocks now matters. One common approach for optimal design with correlated observations is to select the uncorrelated-observation D-optimal design that performs best in the presence of a given correlation structure. In general, however, we see that SA produces a design that is more efficient than this common approach is able to produce.

*Friday, 4:20 pm, Granary Room:*
**Design Aspects for Body Image Measurements**

Craig Johns, (University of Colorado-Denver), cjohns@math.cudenver.edu,
Russel Boice, (UC-Denver), rboice@math.cudenver.edu, and
Rick Gardner, (UC-Denver), Rick.Gardner@cudenver.edu

Self-perception of body image has been shown to have a link with several forms of pathology including eating disorders, depression and psychological distress. Recently, an automated computer program was released which uses readily available digital photography equipment to record psychophysical measurements of a subject's self-perception of body image. The interactive nature of the program leads to questions of statistical design. We discuss statistical models which incorporate the psychophysical data collection as well as design issues for dependent Bernoulli trials.

*Friday, 4:40 pm, Granary Room:*
**A Stabilized Lugannani-Rice Formula**

George Terrell, (Virginia Polytechnic Institute), terrell@vt.edu

The well-known approximate c.d.f formula proposed by Lugannani and Rice (1980) provides in many cases remarkably better accuracy than naive normal approximations. In extreme cases, however, it need not even be a probability. We propose a modification of the formula that is of the same asymptotic order of accuracy and similar typical performance, but is a stable approximation. That is, it is always a probability; and it is exact on an important family of arbitrarily skewed random variables.

*Friday, 5:00 pm, Granary Room:*
**Simulation from a Normally Weighted Dirichlet Distribution**

Alan Genz, (Washington State University), alangenz@wsu.edu

Analysis of a class of non-neutral population genetics models requires the simulation of samples, computation of sample statistics and evaluation of likelihood surfaces for a multivariate density function that is the product of a Normal density and a Dirichlet density, defined over an m-dimensional simplex. A method will be described for the accurate computation of the normalizing constant for the distribution function. Several methods will be considered for computing samples from the distribution. The efficiency of the methods will be compared using results from application examples with as many as twenty-five variables.

## Refereed Contributed Papers II: Nonparametrics

(Refereed Contributed Session)

Session Chair: Randy Eubank

*Friday, 4:00 pm, Market Street Room:*
**Estimating Partially Linear Models Using Wavelets: A Nonlinear Backing Algorithm**

Leming Qu, (Boise State University), qu@math.boisestate.edu

Partially linear models have a linear part as in the linear regression and a nonlinear part similar to that in the nonparametric regression. The estimates in partially linear models have been studied previously in traditional smoothing methods such as smoothing spline, kernel and piecewise polynomial methods. In this paper, we apply the regularized wavelet estimators by penalizing the $l_1$ norm of the wavelet coefficients of the nonparametric function. The regularization parameter is chosen by universal threshold. When the linear part has multivariate predictors, we developed an iterative algorithm similar to backfitting based on the necessary and sufficient conditions of the minimum point. Simulation results confirmed the good performance of the regularized wavelet approach.

*Friday, 4:25 pm, Market Street Room:*
**A Two-Dimensional Robust Nonlinear Smoother for Irregularly Spaced Data**

Karen Kafadar, (University of Colorado at Denver), kk@math.cudenver.edu, and
Max Morris, (Iowa State University), mmorris@iastate.edu

Tukey and Tukey (1982) proposed a nonlinear two-dimensional smoother that is based on medians and thus is highly robust. The motivation for this smoother was irregularly spaced, possibly sparse, data in the plane, such as mortality rates or environmental data reported in census tracts or counties. While the Tukeys' "headbanging" smoother performs well with appropriately selected parameters, it can have the undesirable effect of providing a smoothed value from perfectly co-planar data points that itself does not lie in the plane; i.e., it may fail to pass linear surfaces. We propose a nonlinear smoother that is inspired by headbanging and shares many of its properties, but which leaves entirely unaltered any data set whose data points fall along a (noiseless) first-order polynomial in $(x, y)$. We demonstrate its performance via simulation, discuss some computational challenges in the algorithm, and present an example using temperature trends at various monitoring stations across the United States.

*Friday, 4:50 pm, Market Street Room:*

### A Comparison of Filter and Wrapper Methods for Feature Selection in Supervised Classification

Edgar Acuna, (University of Puerto Rico at Mayaguez), edgar@cs.uprm.edu

In this paper we carry out an empirical comparison of the performance of filter and wrapper procedures for feature selection in supervised classification. The filter methods considered are the RELIEF, Las Vegas Filter, and a new procedure that is being introduced here called FINCO. Among the wrapper methods we considered sequential forward selection, sequential backward selection and the sequential floating forward selection. The classifier used for the wrapper methods is one based on kernel density estimation. Both type of procedures are compared according to their percentages of features selected and their effect in the misclassification error rate of a kernel density estimate classifier. The comparison is carried out in twelve datasets coming from the Machine Learning Database Repository at the University of California, Irvine.

*Friday, 5:15 pm, Market Street Room:*

### Novel Methods for Multivariate Ordinal Data applied to Genetic Haplotypes, Genomic Pathways, Risk Profiles, and Pattern Similarity

Knut M. Wittkowski, (The Rockefeller University), kmw@rockefeller.edu

Introduction: Conventional statistical methods for multivariate data (e.g., discriminant/ regression) are based on the (generalized) linear model, i.e., the data are interpreted as points in a Euclidian space of independent dimensions. The dimensionality of the data is then reduced by assuming the components to be related by a specific function of known type (linear, exponential, etc.), which allows the distance of each point from a hyperspace to be determined. While mathematically elegant, these approaches may have shortcomings when applied to real world applications where the relative importance, the functional relationship, and the correlation among the variables tend to be un-known. Still, in many applications, each variable can be assumed to have at least an "orientation", i.e., it can rea-sonably assumed that, if all other conditions are held constant, an increase in this variable is either "good" or "bad". The direction of this orientation can be known or unknown. In genetics, for instance, having more "abnormal" alleles may increase the risk (or magnitude) of a disease pheno-type. In genomics, the expression of several related genes may indicate disease activity. When screening for security risks, more indicators for atypical behavior may constitute raise more concern, in face or voice recognition, more indicators being similar may increase the likelihood of a person being identified. Methods: In 1998, we developed a non-parametric method for analyzing multivariate ordinal data to assess the overall risk of HIV infection based on different types of behavior[1] or the overall protective effect of barrier meth-ods against HIV infection.[2] By using u-statistics, rather than the marginal likelihood, we were able to increase the computational efficiency of this approach by several orders of magnitude. Results: We applied this approach to assessing immunogenicity of a vaccination strategy in cancer patients.[3] While discussing the pitfalls of the conventional methods for linking quantitative traits to haplotypes, we realized that this approach could be easily modified into to a statistically valid alternative to a previously proposed ap-proaches.[4] We have now begun to use the same methodology to correlate activity of anti-inflammatory drugs along genomic pathways with disease severity of psoriasis based on several clinical and histological characteristics. Conclusion: Multivariate ordinal data are frequently observed to assess semi-quantitative characteristics, such as risk profiles (genetic, genomic, or security) or similarity of pattern (faces, voices, behaviors). The conventional methods require empirical validation, because the functions and weights chosen cannot be justified on theoretical grounds. The proposed statistical method for analyzing profiles of ordinal variables, is intrinsically valid. Since no additional assumptions need to be made, the often time-consuming empirical validation can be skipped.

## Data Management for Statistical Data Bases

(Invited Session)

Organizer and Session Chair: Andrew Westlake

*Saturday, 8:00 am, Weights & Measures Room:*
### Database Technology for Statistical Data

Arie Shoshani, (Lawrence Berkeley Laboratory), arie@lbl.gov

Most statistical analysis is performed by statistical packages such as SAS or R. However, these systems offer little help in managing data before they are analyzed. Rather, it is expected that the data are organized ahead of time in a format suitable for the statistical package. While relational databases can be used to store statistical data, the table structure of the relational model is not the most appropriate for statistical data. In this talk, we will discuss advances in database technology designed for managing and manipulating statistical data. In particular, the areas of statistical databases and OLAP (On-Line Analytical Processing) have developed models suitable to represent multi-dimensional data where each dimension can be further organized as a hierarchy of categorical attributes. Further, some database products were built to optimize queries over such data. We will also discuss the concept of federating statistical databases with ordinary object databases, which typically contain other data or metadata associated with the statistical data. The purpose of the federation is to provide a capability of jointly querying the databases in the federation while maintaining their independence. We will present a prototype system that supports such a federation.

*Saturday, 8:35 am, Weights & Measures Room:*
### Metadata Usage in Statistical Computing

Wilfried Grossmann, (Dept. of Statistics and Decision Support Systems, University Vienna), wilfried.grossmann@univie.ac.at

Information about data plays a crucial role in all steps of statistical analysis, but (data) management of this information is usual done in an ad-hoc manner by the working statistician. Recent developments of statistical computing environments, in particular for data mining, have improved the situation, but a systematic approach is not yet available. In the talk we will outline a model, which integrates statistical data and information about the data (i.e. metadata), and show the application of this structure in the context of statistical computing. The model is based on a number of information objects describing the data in some detail, for example information about the underlying populations, the methods for obtaining the data, or the variables used in the data set together with their roles in the context of the data set. Based on this model one can define for each statistical procedure the corresponding transformations on the adjoined metadata and describe the modifications of the metadata implied by the statistical procedure. Examples for a number of important data pre-processing steps like data combination, modification of variables or weighting is given.

*Saturday, 9:10 am, Weights & Measures Room:*
**Data Structures for HIV and AIDS Notification and Analysis**

Andrew Westlake, (Survey & Statistical Computing), ajw@sasc.co.uk

Most statistical procedures are based on a single rectangular dataset, and much statistical data is con-ceived and stored in this form. Database professionals usually start from the relational database model for their data structure designs, effectively a set of related rectangular datasets. This allows the accurate representation of more complex structures, and, where this complexity is relevant, allows easier mainte-nance and manipulation of the data. This paper describes the re-structuring of the processing and analysis system for the HIV and AIDS reporting system for England and Wales, and discusses the benefits that were achieved, in terms of easier processing and maintenance, more flexible manipulation, and faster reporting.

## Surveillance

(Invited Session)

Organizer and Session Chair: Edward J. Wegman

*Saturday, 8:00 am, Smokehouse/Rap of the Gavel Room:*
**Multiscale 'Spatial' Analysis of Network Data: Putting Wavelets on Graphs.**

Eric D. Kolaczyk, (Boston University), kolaczyk@math.bu.edu

Massive amounts of data currently are being gathered in a variety of contexts in which the underlying structure is that of a network. This includes measurements of intra- and internet traffic, but also data from sensor networks, social networks, etc. Often, though not always, there is a well-defined spatial context corresponding to the network, such as that induced by router or sensor locations. However, strict interpretation of the data within this context can be misleading, and viewing the measurements from the context of a graph topology typically is more appropriate. Nevertheless, just as in traditional spatial analysis, the concept of 'scale' can play an important role in the analysis of network data, such as in problems centered on the detection of anomalies or the determination of the concentration/diffuseness of an attack. I will present an overview of ongoing work with colleagues on the development of tools for the analysis of graph-indexed data at multiple scales. A common theme throughout will be on extensions of the wavelet paradigm to arbitrary graphs, as well as descriptive and inferential tools for analysis under the resulting alternative data representation. Illustrations will be provided using internet traffic data.

*Saturday, 8:35 am, Smokehouse/Rap of the Gavel Room:*
**Social Networks and Computer Networks**

John Rigsby, (Naval Surface Warfare Center Dahlgren Division), rigsbyjt@nswc.navy.mil,
and
Jeff Solka, (Naval Surface Warfare Center Dahlgren Division), solkajl@nswc.navy.mil

Understanding how network infrastructure changes with time is essential to protecting an organization's network. Multiple methods for discovering network topology using different areas of graph theory and concepts of social network relationships will be discussed. Passively detecting changes in network topology and presenting this to network engineers and analysts will increase an organization's threat management capabilities to counter malicious network activities; the application of change point detection to social networks will be the backbone of this approach. This research project is in the early stages of development and will be presented as such.

*Saturday, 9:10 am, Smokehouse/Rap of the Gavel Room:*
**Classification Complexity Measures and Their Relationship to Feature Selection**

Jeff Solka, (NSWCDD), solkajl@nswc.navy.mil, and
David Johannsen, (NSWCDD), johannsendj@nswc.navy.mil

This talk will explore some of our recent work in the use of new classification complexity measures as surrogates to actual nearest neighbor performance. The talk will discuss the relationship between these surrogate measures and the actual cross validated nearest neighbor performance. We will also provide discussions on the use of these measures in schemes for classifier optimization under Minkowski metric space adaptation.

## Prediction of Catastrophic Events

(Invited Session)

Organizer and Session Chair: Amy Braverman

*Saturday, 8:00 am, Granary Room:*
**Detecting Features in Seismic and Geodetic Data**

Andrea Donnellan, (Jet Propulsion Laboratory), donnellan@jpl.nasa.gov,
Robert Granat, (Jet Propulsion Laboratory), Robert.Granat@jpl.nasa.gov, and
John Rundle, (University of California, Davis), jbrundle@ucdavis.edu

Earthquakes are non-regular in both time and space making forecasting of future earthquakes difficult. Additionally they occur on timescales of hundreds to thousands of years. Seismic data and more recently geodetic data are used to study earthquake processes as well as geological study of earthquake faults. Earthquake faults are part of larger systems in which faults interact with each other. An earthquake on one fault can either inhibit or increase stress on nearby faults. We are using multiple approaches to better understand the earthquake processes and interacting fault systems. These include simulating earthquake processes as well as looking for patterns in seismic and geodetic data. We are using extensions of principle component analysis and hidden Markov modeling to detect features and anomalies in the data. The approaches show promise toward forecasting future earthquakes.

*Saturday, 8:35 am, Granary Room:*
**Predicting Damaging Climate Events: Methods, Examples, and Public Reaction**

David W. Pierce, (Scripps Inst. Oceanography, Climate Research Division),
dpierce@ucsd.edu, and
Tim P. Barnett, (Scripps Inst. Oceanography, Climate Research Division), tbarnett@ucsd.edu

Climate events can cause significant damage to parts of the United States or world. The likelihood of occurrence for some damaging events are predictable months to years in advance, in cases where the forcing and physics of the phenomena are sufficiently well understood. Two examples of this will be presented. First, the massive El Nino of 1997/98 was predicted a year in advance, and generated intense public interest. This El Nino resulted in a significant amount of damage over portions of the United States, but also less severe than usual winter conditions over other parts. Our experience was that conveying the probabilistic nature of such events to the general public was difficult, and the results were often misinterpreted. Despite this, there is evidence that mitigation efforts (presumably spurred by the widespread media coverage) reduced damage from this event by a significant amount. Second, the Accelerated Climate Prediction Initiative (ACPI) project has used computer models of global warming to determine the changes in available water resources over the western United States for the next fifty years. The results suggest a high likelihood of large changes in the yearly water cycle, whose effects will be disastrous if not mitigated. The probabilistic nature of such forecasts will be discussed, and the reaction to these forecasts by state government authorities will be outlined.

*Saturday, 9:10 am, Granary Room:*
**Predicting and Comprehending Asteroid Impacts**

Clark R. Chapman, (Southwest Research Inst.), cchapman@boulder.swri.edu

Asteroids and comets have struck our planet, with devastating consequences, during geological history. Very improbably during our lifetimes, one may strike again with sufficient energy to destroy civilization as we know it or – even more improbably – erase the human species from planet Earth. This "impact hazard" has a number of unusual features that should interest experts in statistics. As an astronomer, not a mathematician, I discuss several of these issues.

First, the prediction of impacts might, at first glance, seem to be a "particle-in-a-box" theoretical problem, and then a straightforward "missile impact" problem when an asteroid is detected that is heading our way. But the reality is different. The orbital dynamical aspects of the problem, in the context of the telescopic observations that discover and track the potential impactors, are highly non-intuitive.

Second, the impact hazard challenges ordinary citizens, and more scientifically educated people as well, to grabble with the essence of very low probability but high consequence events. It is like the lottery in reverse. Whether society should take this hazard seriously, and invest in practical technological approaches to mitigating the hazard, depends on effective communication between technical experts and the lay public and decision makers who set national priorities. Especially in the context of 9/11 terrorism, an understanding of how ordinary people relate to statistics in the context of potential danger is a central issue for our society.

## Graphics for Bio and Chem Informatics

(Invited Session)

Organizer and Session Chair: Daniel B. Carr

*Saturday, 8:00 am, Market Street Room:*
**Grapic-Centric, Computationally-Efficient Recursive Partitioning**

James Vivian, (Golden Helix, Inc., Bozeman, MT), vivian@goldenhelix.com,
S. Stan Young, (CGStat LLC, Raleigh, NC), genetree@bellsouth.net, and
Christophe Lambert, (Golden Helix, Inc., Bozeman, MT), lambert@goldenhelix.com

FIRM (Formal Inference-based Recursive Modeling) is a hypothesis testing-based, multiway splitting version of recursive partitioning that offers significan advantages in computational speed and intrepretability. We present a GUI incarnation of FIRM which is interactive for modestly large data sets, 10k observations and 1k descriptors, and provides decision-assisting graphics. There are applications, one customized for the analysis of high-throughput screening data in drug discovery, and the other for pharmacogenetics analysis, i.e. linking patient drug-response to genotype/phenotype information. The visual nature of the recursive partitioning 'trees' and the usual availability of alternative split variable both point to the need for interactive decision-making for tree building. The chemistry software builds chemically intuitive descriptors and marks important split features in molecules. The pharmacogentics application builds in complex statistical genetics methods: linkage disequilibrium analysis, visualization of e.g. Hardy-Weinberg equilibrium, and a tractable (and ruthlessly efficient) method to impute haplotypes. Using examples from public datasets, we will demonstrate the power, efficiency, and intuitive appeal of a GUI recursive-partitioning to unravel complex patterns and relationships in large datasets leading to the transformation of data to information. We will emphasize the integration of graphics into a user-centric recursive partitioning analysis.

*Saturday, 8:45 am, Market Street Room:*

### Applications of Computational Geometry, Statistical Analysis, and Graphics to the Study of Molecular Systems

Daniel B. Carr, (George Mason University), dcarr@gmu.edu, and
Iosif Vaisman, (George Mason University), ivaisman@gmu.edu

Methods of computational geometry provide a robust and effective approach to studying topology and architecture of molecular systems. A molecule or molecular system can be represented by the set of points in three-dimensional space, where each point designate an atom or a site (group of atoms). The Delaunay tessellation of such a set of points generates an aggregate of space-filling irregular tetrahedra, or Delaunay simplices. The vertices of each simplex define objectively four nearest neighbor atoms. The collection of all simplices describes the topology of a molecular system. We use statistical analysis of geometrical and compositional properties of the Delaunay simplices to characterize structure and connectivity of the molecular system and to correlate chemical composition with the three-dimensional molecular architecture. In protein structure analysis the Delaunay tessellation facilitates objective identification of neighboring residues for a quantitative description of nonlocal contacts in three-dimensional protein structures. Analysis of the patterns of spatial proximity of residues in known protein structures based on the Delaunay tessellation reveals highly nonrandom clustering of amino acids. The talk presents a variety of graphics for showing statistics associated with the molecular systems. We also present graphics showing statistics computed from databases of peptides that bind with immune system molecules.

## Statistical Methods to Compress and Query Large Databases and Data Streams

(Invited Session)

Organizer and Session Chair: Silvia Nittel

*Saturday, 10:30 am, Weights & Measures Room:*

### Data Stream Algorithmics

S. Muthukrishnan, (AT&T and Rutgers Univ), muthu@cs.rutgers.edu

Say Paul shows numbers 1 through N in some permuted order, but leaves out one (or two) of the numbers. Carole needs to determine the missing number(s). But Carole has the ability to memorize only a small set of the numbers she sees! As one would expect, it suffices for Carole to simply keep certain statistics on the input stream to solve this problem.

This simple puzzle is an example of processing data streams. In many applications such as in IP traffic analysis and in Homeland Security, data arrives as a stream. Keeping certain statistics on the data stream helps solve a number of monitoring problems. The challenge is to design suitable algorithms for estimating these statistics within limited resources.

I will present an idiosyncratic view of this emerging area of data stream algorithmics. I will also discuss how to abstract these solutions and concerns into data stream management systems.

*Saturday, 11:05 am, Weights & Measures Room:*

### Efficient Processing of Massive Data Streams for Mining and Monitoring

Mirek Riedewald, (Cornell University), mirek@cs.cornell.edu,
Johannes Gehrke, (Cornell University), johannes@cs.cornell.edu,
Alan Demers, (Cornell University), ademers@cs.cornell.edu,
Abhinandan Das, (Cornell University), asdas@cs.cornell.edu, and
Alin Dobra, (Cornell University), dobra@cs.cornell.edu

Data arriving as high-speed data streams poses a serious challenge for data management as the traditional DBMS paradigm of set-oriented processing of disk-resident records does not apply. Especially problematic are blocking operators and operators with unbounded state for infinite input, because their memory footprint might grow without bounds. At Cornell University we are designing a system for distributed data stream mining and monitoring. In this talk, we will first overview some open research challenges in processing data streams, and then we will describe algorithms for the approximate computation of set-valued query results.

*Saturday, 11:400 am, Weights & Measures Room:*
**Wavelets for Efficient Querying of Large Multidimensional Data Sets**

Cyrus Shahabi, (USC), shahabi@usc.edu

New data-intensive applications operate on diverse types of data, signifying new characteristics in both querying the data and obtaining the results. In particular: 1) the data set is large and multidimensional; popular examples are spatial and temporal data, as well as sensor data streams, 2) the queries are complex, asking for trends or outliers in data, correlation between different dimensions, or aggregation of one or more (measure) attributes given a bounded multidimensional space (termed, range-aggregate queries), and 3) the results can be approximate and/or progressively become exact. These characteristics lead us to believe that wavelet transform will become a likely tool for future database query processing. Although a straightforward adoption of wavelet is to utilize it to reduce the data size at different resolutions, unfortunately data compression methods are only effective on datasets that compress well and on queries that require the reconstruction of the entire signal. Therefore, we are taking a fundamentally different approach. Our approach is a data independent approximation technique that is based on query approximation rather than data compression. Many common queries, including relational algebra expressions and a large class of aggregate queries, can be conceptualized as performing a linear transformation on the density distribution of an input relation to produce the density of an output relation. This observation leads us to the following generic progressive query evaluation plan: decompose the query transformation into a series of small, cheaply computable sub-transformations and evaluate the most important sub-transformations first. In this paper, we discuss the details of this evaluation plan for polynomial range-sum queries, and show how it can be extended to support a batch of several submitted queries. In addition, we show that typical queries on wavelet data require a distinct access pattern which we describe and exploit to design a disk placement strategy for wavelet data that yields best-possible I/O complexity for point and range query evaluation. We conclude by discussing some open problems in dealing with wavelets from a database perspective such as how to perform I/O efficient wavelet transformation.

## Data Mining Combat Simulations

(Invited Session)

Organizer and Session Chair: Barry A. Bodt

*Saturday, 10:30 am, Smokehouse/Rap of the Gavel Room:*
**Data Mining Combat Simulations: an Emerging Opportunity**

Barry A. Bodt, (U.S. Army Research Laboratory), babodt@arl.army.mil

We stand at an interesting technology crossroads. Data mining methods, high performance computing, and high fidelity stochastic simulations are now sufficiently mature to potentially bring about great improvement in the manner in which a battle commander plans and executes a battle. In this presentation, I will set the stage for the other speakers in this session by outlining briefly what has been attained recently in Army computing and simulation and argue that the product is worthy of data mining. I will contrast how Army battle planning is performed now with how it might be in the future given assistance from data mining and also suggest practical limits on expectations for implementation. Finally, I will discuss the specific stochastic simulation model and scenario used in this pilot study to generate the data mined by the following speakers.

*Saturday, 10:55 am, Smokehouse/Rap of the Gavel Room:*
**Regression Tree Analysis of Battle Simulation Data**

Wei–Yin Loh, (University of Wisconsin, Madison), loh@cs.wisc.edu

The sparse and high-dimensional nature of data from battle simulations creates two difficult problems. First, many regression methods cannot deal with these data without some sort of variable selection. If the wrong variables are deleted, much information can be lost. Second, the models are usually difficult to interpret. Even when interpretation is possible, the conclusions can be wrong.

We introduce some new ways of thinking about regression trees that can overcome the above problems. The new methods produce models that are instantly recognizable, are not biased by variable selection, and have high prediction accuracy relative to the best tree and non-tree methods.

*Saturday, 11:20 am, Smokehouse/Rap of the Gavel Room:*
**Robust Modeling Based on L2E Applied to Combat Simulation Data**

David Kim, (United States Military Academy), ad1384@exmail.usma.army.mil

Parametric modeling based on a minimum distance criterion gives us a robust(i.e. resistant to outliers and other types of data contamination) way of analyzing the data. The integrated squared error (L2E) of David Scott is one such criterion, and a system of fitting a model to the data and assessing the model can be built upon it. Combat simulation data containing numerous variables may benefit from such a robust method of analysis since it is quite likely that some of the assumptions for building, for example, a least squares linear model may be violated in the data.

We will outline the development of such a system and demonstrate the new methodologies via the analysis of simulation results from One Semi Automated Forces (OneSAF). Specifically, the variable selection methods based on L2E are used in the analysis to provide models identifying key battle parameters for a given engagement.

*Saturday, 11:45 am, Smokehouse/Rap of the Gavel Room:*
**Discovery of Battle States Knowledge from Muti-Dimensional Time Series Data**

T. W. Liao, (Louisianna State University), ieliao@lsu.edu,
B. Bodt, (U.S. Army Research Laboratory), babodt@arl.army.mil,
J. Forester, (U.S. Army Research Laboratory), forester@arl.army.mil,
C. Hansen, (U.S. Army Research Laboratory), charlie@arl.army.mil,
E. Heilman, (U.S. Army Research Laboratory), heilman@arl.army.mil,
C. Kaste, (U.S. Army Research Laboratory), rck@arl.army.mil, and
J. O'May, (U.S. Army Research Laboratory), jomay@arl.army.mil

This paper presents a suite of knowledge discovery techniques developed for processing multi-dimensional battlespace time series data in order to capture the knowledge of determining battle states in the form of fuzzy rules. This discovered knowledge is subsequently applied to provide the battle staff (or the nerve center of the Future Combat System) with current battle state information and to predict future battle outcomes. The major techniques employed include interpolation for preprocessing data, clustering to establish the concept of battle states, and genetic-fuzzy modeling for time series forecasting and classification. The clustering results are first checked by human experts and then used as the input data for the discovery of battle state determination. Two knowledge discovery methods have been implemented: WM-based genetic-fuzzy and LIAO-based genetic fuzzy. Represented in the form of fuzzy rules, the discovered knowledge is very easy to comprehend if the number of fuzzy terms is not too high. The predicted data are fed into the knowledge of determining battle states to predict future battle states. This is easily achieved by a fuzzy reasoning method such as max-min.

Data are needed to test the effectiveness of this tool. To this end, we created a battle scenario using the OneSAF combat simulation. In the full paper, more details will be given to describe the experiments and the Killer-Victim Scoreboard (KVS) data collection method. The discovered knowledge and their use will also be presented as well.

## Forests at Risk

(Invited Session)

Organizer and Session Chair: Gretchen Moisen

*Saturday, 10:30 am, Granary Room:*

### Satellite to the Public in Near Real-Time: Providing Active Wildfire Information with MODIS Rapid Response

Mark Finco, (RedCastle Resources / USDA Forest Service), mfinco@fs.fed.us,
Brad Quayle, (USDA Forest Service), bquayle@fs.fed.us,
Rob Sohlberg, (University of Maryland), rsohlber@geog.umd.edu, and
Jacques Descloitres, (Goddard Space Flight Center), jack@ltpmail.gsfc.nasa.gov

Over the past several years, severe wildfire seasons have become the norm in the United States. When many large wildfires burn simultaneously, fire fighting assets often need to be shared between geographic areas of the country. The MODIS Rapid Response (RR) system is a collaborative effort between the US Forest Service (USFS), NASA's Goddard Space Flight Center, and the University of Maryland, which delivers near real-time active fire locations to wildfire managers. Twice daily, RR provides maps of active fires over the contiguous United States and Alaska. This "big picture" with respect to spatial extent and juxtaposition of wildfires is an important information asset for the allocation of finite fire fighting resources. This paper describes the heterogeneous computing environment that is RR, including how multiple sources of satellite data are used synergistically and how this data is delivered as maps on the Internet.

*Saturday, 11:05 am, Granary Room:*

### Identifying "Redtops": Classification of Satellite Imagery for Tracking Mountain Pine Beetle Progression through a Pine Forest

Richard Cutler, (Department of Mathematics and Statistics, Utah State University), richard@math.usu.edu,
James Powell, (Department of Mathematics and Statistics, USU), powell@math.usu.edu,
Leslie Brown, (USDA FS Rocky Mountain Research Station), lbrown@cc.usu.edu, and
Barbara Bentz, (USDA FS Rocky Mountain Research Station), bbentz@fs.fed.us

Mountain pine beetles are an integral part of moutain pine forest ecosystems in the intermountain west. Locally, mountain pine beetle attacks may be devastating. With global warming comes the concern that mountain pine beetles may soon be able to attack different species of trees at higher elevations that have no resistance to the beetles.

A collaborative project, between members of the Department of Mathematics and Statistics at Utah State University, and the USDA Forest Service Rocky Mountain Research Station has focused on the use of differential equation models for the mountain pine beetle life cycle and attack mechanism. The work reported here concerns the identification of trees attacked by mountain pine beetles and, more generally, of the vegetation in the Sawtooth National Forest using satellite imagery and ground observations. The satellite imagery has 4m (4 bands) and 1m (panchromatic) resolution. Classification methods applied to the data include CARTs, random forests, and more traditional methods. Results for the different classifiers are presented, along with a discussion of issues concering the matching of ground locations to satellite image pixels.

*Saturday, 11:40 am, Granary Room:*

### Design Attributes for Sampling Rare Ecological Events in Forest Ecosystems: Lichens in the Pacific Northwest

Thomas C. Edwards, (US Geological Survey, Utah Cooperative Research Unit, Utah State University), tce@nr.usu.edu, and
Richard Cutler, (Dept. of Mathematics and Statistics, USU), richard@sunfs.math.usu.edu

Despite the fact that sampling for rare events would seem to be a common problem in ecology, there is surprisingly little literature on the subject. This paper describes, and discusses the advantages and disadvantages of, a number of the most commonly used probability and non-random sampling methods for sampling rare ecological events in forest ecosystems. Focus is on a suite of species in the US Pacific Northwest referred to as Survey and Manage species, and includes fungi, lichens, bryophytes, mollusks, a few vascular plants, arthropods and two vertebrates. All represent species for which little information is known, and this lack of knowledge make their management quite difficult. Sampling methods considered include stratified random sampling with disproportionate allocation, network and snowball sampling, adaptive methods such as adaptive cluster sampling, and model-based ppz-sampling. Information needs of the US Forest Service and logistical concerns that constrained the eventual choice of sample designs are discussed. The probability sampling framework that was finally implemented for the Survey and Manage program, and how it addresses information needs for forest management, are described in detail. Some important elements of the framework are the use of an existing vegetation survey grid for sampling locations, and stratification (with disproportionate allocation) by stand age class and late successional reserve status. The paper concludes with some recommendations for future probability sampling efforts for the Survey and Manage program.

## Nonparametrics Methods and Applications

(Contributed Session)

Session Chair: Zewen Zhu

*Saturday, 10:30 am, Market Street Room:*

### Multivariate Density Estimation with Permuted Variable-Values

Sridevi Parise, (University of California, Irvine), sparise@ics.uci.edu,
Padhraic Smyth, (University of California, Irvine), smyth@ics.uci.edu, and
Sergey Kirshner, (University of California, Irvine), skirshne@ics.uci.edu

In this paper we consider the problem of multivariate density estimation from a data set of N rows (observations) and p columns (variables), where in each row the observed values can be permuted in some random fashion. Thus, for any row we do not know with certainty which variable is associated with each value. This problem arises in practice in a number of practical applications such as computer vision (where measured features of objects are not in correspondence with each other) and in text analysis (where fields of the data are not in correspondence). We derive lower-bounds on how well any algorithm can "un-mix" the rows. The concepts in the paper are illustrated using an application in astronomical data analysis, involving registration, clustering, and classification of images of galaxies.

*Saturday, 10:50 am, Market Street Room:*

### Computational Challenges in Computing Nearest Neighbor Estimates of Entropy for Large Molecules

E. James Harner, (West Virginia University), jharner@stat.wvu.edu,
Harshinder Singh, (NIOSH/West Virginia University), hsingh@stat.wvu.edu,
Shengqiao Li, (NIOSH/West Virginia University), shli@stat.wvu.edu, and
Jun Tan, (West Virginia University), jtan@stat.wvu.edu

Entropy is a statistical thermodynamic property of molecules; its evaluation is important for studying the properties of biological molecules (such as peptides, proteins, and DNA molecules) and chemical molecules. Entropy evaluation is also important in drug designs and for investigating the effect of toxins on human skin. The entropy of a molecule depends mainly on random fluctuations in its torsional (also called rotational or dihedral) angles. The traditional approach assumed a multivariate normal distribution for the torsional angles of large molecules (Karplus and Kushik, Macromolecules, 1981). However, the assumption of normality is not valid in many situations, particularly when there are large fluctuations in the torsional angles. Demchuk and Singh (Molecular Physics, 2001) introduced a circular probability approach to modeling torsional angles in molecules and illustrated the modeling of the torsional angle of methanol using von Mises distributions. A bathtub shaped probability distribution was derived for the potential energy of the methanol molecule. Singh et al (Biometrika, 2002) introduced a bivariate circular model, which is a natural torus version of the bivariate normal distribution to which it reduces when the fluctuations in the angles are small. The marginal distributions are symmetric and are either unimodal or bimodal. This model was used for modeling two angels of a pentapeptide. In general, the torsional angles can have arbitrary shapes and macromolecules have a large number of torsional angles, which are interdependent. Thus a nonparametric approach appears to be a natural choice for entropy estimation of large molecules. However, entropy evaluation using histogram and kernel density estimates also has problems in high dimensions.

Estimates of entropy based on nearest neighbor distances between sample points (Kozachenko and Leonenko. Problems of Information Transmission, 1987) and estimates of entropy based on kth nearest neighbor distances (Singh et al., to appear in Statistics and Decisions) offer hope for estimating entropy for large molecules. However, evaluating nearest neighbor distances is computationally challenging when the number of torsional angles is large and the data obtained using molecular dynamic simulations on the molecule is huge. We discuss computational approaches for obtaining estimates of entropy based on nearest neighbor distances. Our approaches use RMPI to parallelize $n^2$ and $n\log(n)$ kth nearest neighbor algorithms, where n is the number of dynamic simulations. We illustrate this approach using data on torsional angles of some large molecules.

*Saturday, 11:10 am, Market Street Room:*

### Mixture Transitions for Edge Preservation in Kalman Filtering

Mark Fitzgerald, (University of Colorado, Denver), fitz@math.cudenver.edu

In certain applications, it is desirable to fit a smooth curve to a set of (possibly multivariate) observations while preserving discontinuities where they exist. The standard Kalman filter provides a simple mechanism for estimating a smooth sequence from equally-spaced observations, but it will over-smooth discontinuities. In the standard Kalman filter, it is assumed that the difference between consecutive observations has small variance (relative to the noise). An extension of that model allows for the difference to come from a mixture of two distributions, one with small and one with large variance. The resulting estimator provides a smooth sequence, but with large jumps where the data indicate the presence of discontinuities.

*Saturday, 11:30 am, Market Street Room:*

**Statistical Learning Theory and Statistics: Embracing New Technologies**

Kevin Watanabe, (Kxen, Inc.), kevin.watanabe@kxen.com

In an article by J. Friedman, he made a commentary on the prevailing mindset within the statistics profession. He stated that "Most statisticians seem to agree that statistics is becoming relatively less influential among the information sciences." He further suggested, "Perhaps more than at anytime in the past statistics is at a crossroads: we can decide to accommodate or resist change." He further added "...we may have to moderate our tendency to disregard developments...that appear to work well, simply because the reasons for their success are not yet well understood by us."

Statistical Learning Theory (SLT) is one such method developed by a Russian mathematician named Vladimir Vapnik. A recent popular statistical book written by T. Hastie, R. Tibshirani and J. Freidman entitled "The Elements of Statistical Learning" reviewed the current methods within the discipline of learning theory. The methods included expansions, regularization, kernel methods, boosting, neural networks, and support vector machines which is an application of SLT. Many of these methods are embraced by computer scientists but suspiciously viewed by traditional statisticians. I will review current applications of SLT to data preparation, feature space reduction, and regression.

A foundational theory of SLT is structured risk minimization (SRM). The application of SRM theory is to automatically discover the optimal balance between model fit and model consistency. A parameter called VC dimension is used to find this optimal balance. VC dimension is a measure of model complexity but it is not related to the number of model variables as traditional regression methods.

Most of the time required to build a model is spent in data preparation and feature space reduction. SLT can automate this tedious process from days or weeks to minutes or hours. Missing values for continuous variables are replaced with the mean and normalized. In addition, continuous variables are binned relative to the dependent variable while maintaining the natural order. Missing values for categorical variables are automatically grouped into a missing group. Target optimized bins are also created for categorical variables. Therefore variables with high cardinality such as zip code will be binned into a small number of homogeneous groups relative to a target.

Another strength of SLT against other methods such as logistic regression is the ability to build predictive models with all variables in a data warehouse. If the warehouse has 2,000 or more variables, the first predictive model using SLT can model using all variables. The variables will be ordered from the most predictive to the least predictive. Using SLT for exploratory data analysis or data mining is extremely effective in discovering the key drivers very quickly. The "curse of dimensionality" which is a prevailing problem for traditional regression methods is easily overcome using SLT. Models developed automatically in less than a hour using SLT have similar predictive power to tailor made models which can take more than three weeks.

With the tremendous increase in the magnitude and complexity of data being collected daily, new break through technology such as SLT can became a powerful tool for statisticians. The ongoing research of SLT includes the application to unsupervised and supervised clustering and time series. The future of statistics requires the embracing of new technologies such as SLT in order for statistics to keep pace with the emerging technologies being developed by the other information sciences.

*Saturday, 11:50 am, Market Street Room:*

## Using a LOESS Smoother to Estimate the Parameters of an Angular Dependent Distribution of HRR Data

Bradley C. Wallet, (Mission Research Corporation), bwallet@mrcday.com,
Robert W. Hawley, (Mission Research Corporation), rhawley@mrcday.com, and
Troy L. Klein, (Mission Research Corporation), tklein@mrcday.com

High range resolution (HRR) data is an important modality for performing automatic target recognition (ATR) of vehicles. However, the distribution of HRR based features is typically highly dependent upon vehicle azimuth. Typical methods for estimating the associated probability density function (PDF) for a given angle have involved using all observation whose angles are within a given window centered upon the angle of interest. Unfortunately, scarcity of data generally leads to the use of a window that is large relative to the rate of change of the underlying PDF. We present a method by which we used a LOESS smoother to estimate the changing parameters of the distributions of HRR data as a function of angle. We then demonstrate how this point estimate provides better performance than a windowed estimate.

# Author Index

# INTERFACE 2004

# BALTIMORE

**May 26–29, 2004**
**Baltimore Marriott Waterfront**

## THEME: BIOINFORMATICS

**Program Chairs:**

**David J. Marchette**
Naval Surface Warfare Center
marchettedj@nswc.navy.mil

**Jeffrey L. Solka**
Naval Surface Warfare Center
solkajl@nswc.navy.mil

http://www.galaxy.gmu.edu/Interface04

---

# INTERFACE 2005

# ST. LOUIS

**Program Chair:**

**Bill Shannon**
Washington University in St. Louis
School of Medicine
shannon@ilya.wustl.edu