TECHNICAL REPORTS

ROUTING AND ACTION

MEMORANDUM

ROUTING

TO:(1) Mathematical Sciences Division (Arney, David)

Report is available for review

(2) Proposal Files Report No.: Proposal Number: 45810-MA.1

DESCRIPTION OF MATERIAL

CONTRACT OR GRANT NUMBER: W911NF-04-1-0447

INSTITUTION: George Mason University

PRINCIPAL INVESTIGATOR: Edward Wegman

TYPE REPORT: Final Report

DATE RECEIVED: 12/10/2008 12:55:48PM

PERIOD COVERED: 11/1/2004 12:00:00AM through 4/30/2008 12:00:00AM

TITLE: Analytical and Graphical Methods for Streaming Data with Applications to Netcentric Warfare

ACTION TAKEN BY DIVISION

(x) Report has been reviewed for technical sufficiency and IS [x] IS NOT [ ] satisfactory.

(x) Material has been given an OPSEC review and it has been determined to be non sensitive and, except for manuscripts and progress reports, suitable for public release.

(x) Perfomance of the research effort was accomplished in a satisfactory manner and all other technical requirements have been fulfilled.

(x) Based upon my knowledge of the research project, I agree with the patent information disclosed.

Approved by NAE\DAVID.ARNEY1 on 12/10/2008 2:17:14PM

REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
10-12-2008	Final Report	1-Nov-2004 - 30-Apr-2008

4. TITLE AND SUBTITLE

Analytical and Graphical Methods for Streaming Data with Applications to Netcentric Warfare

5a. CONTRACT NUMBER
W911NF-04-1-0447

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Edward J. Wegman

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
George Mason University
Office of Sponsored Programs
4400 University Dr. MSN 4C6
Fairfax, VA 22030 -4444

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
45810-MA.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

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

14. ABSTRACT
This project began with a strong interest in streaming data with particular attention to intrusion detection in computer networks. The netcentric battlefield communications was a particular motivation. Several different techniques for univariate and multivariate probability density estimations were developed with recursive updating. An ability to detect subtle shifts in Internet traffic patterns using streaming Internet headers and the recursive density estimators was demonstrated. We also investigated text streaming data and developed methods for topic identification using mathematical representations of text documents. Finally we have noted the connection between two-mode social network analysis and latent semantic indexing

15. SUBJECT TERMS
streaming data, graph theory, network science, text mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Edward Wegman
U	U	U	SAR		19b. TELEPHONE NUMBER 703-993-1691

REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
10-12-2008	Final Report	1-Nov-2004 - 30-Apr-2008

4. TITLE AND SUBTITLE

Analytical and Graphical Methods for Streaming Data with Applications to Netcentric Warfare

5a. CONTRACT NUMBER
W911NF-04-1-0447

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Edward J. Wegman

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
George Mason University
Office of Sponsored Programs
4400 University Dr. MSN 4C6
Fairfax, VA 22030 -4444

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
45810-MA.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

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

14. ABSTRACT
This project began with a strong interest in streaming data with particular attention to intrusion detection in computer networks. The netcentric battlefield communications was a particular motivation. Several different techniques for univariate and multivariate probability density estimations were developed with recursive updating. An ability to detect subtle shifts in Internet traffic patterns using streaming Internet headers and the recursive density estimators was demonstrated. We also investigated text streaming data and developed methods for topic identification using mathematical representations of text documents. Finally we have noted the connection between two-mode social network analysis and latent semantic indexing

15. SUBJECT TERMS
streaming data, graph theory, network science, text mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Edward Wegman
U	U	U	SAR		19b. TELEPHONE NUMBER 703-993-1691

Report Title

Analytical and Graphical Methods for Streaming Data with Applications to Netcentric Warfare

ABSTRACT

This project began with a strong interest in streaming data with particular attention to intrusion detection in computer networks. The netcentric battlefield communications was a particular motivation. Several different techniques for univariate and multivariate probability density estimations were developed with recursive updating. An ability to detect subtle shifts in Internet traffic patterns using streaming Internet headers and the recursive density estimators was demonstrated. We also investigated text streaming data and developed methods for topic identification using mathematical representations of text documents. Finally we have noted the connection between two-mode social network analysis and latent semantic indexing using term-document and bigram-document matrices.

List of papers submitted or published that acknowledge ARO support during this reporting period. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

1. Solka, Jeffrey L., Wegman, Edward J., and Marchette, David J. (2004) "Data mining strategies for detection of chemical warfare agents," Statistical Data Mining and Knowledge Discovery, 71-92

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

3. Johannsen, D.A., Wegman, E.J., Solka, J.L. and Priebe, C.E. (2004) "Simultaneous selection of features and metric for optimal nearest neighbor classification," Communications in Statistics: Theory and Methods, 2137-2158

4. Kafadar, Karen and Wegman, Edward J. (2006) "Visualizing 'typical' and 'exotic' Internet traffic data," Computational Statistics and Data Analysis, 50(12), 3721-3743

5. Dorfman, Alan H., Lent, Janice, Leaver, Sylvia G. and Wegman, Edward J. (2006) "On sample survey designs for consumer price indexes," Survey Methodology, 32(2), 197-216

6. Said, Yasmin H., Wegman, Edward J., Sharabati, Walid K. and Rigsby, John T. (2008) "Style of author-coauthor social networks," Computational Statistics and Data Analysis, 52, 2177-2184, 2008; doi:10.1016/j.csda.2007.07.021, 2007

7. Said, Yasmin H. and Wegman, Edward J. (2007) "Quantitative assessments of alcohol-related outcomes," Chance, 20(3), 17-25

8. Wegman, Edward J. and Martinez, Wendy L. (2007) "A conversation with Dorothy Gilford," Statistical Science, 22(2), 291-300

9. Said, Yasmin H. and Wegman, Edward J. (2008) "Using administrative data to estimate cyclic effects of alcohol usage (refereed abstract)," Alcoholism: Clinical and Experimental Research, 32(6) Supplement, 139A

10. Wegman, Edward J. and Said, Yasmin H. (2008) "Modeling spatiotemporal effects for acute outcomes in an alcohol system (refereed abstract)," Alcoholism: Clinical and Experimental Research, 32(6) Supplement, 140A, 2008

11. Reyen, Salem S., Miller, John J. and Wegman, Edward J. (2008) "Separating a mixture of two normals with proportional covariances," Metrika, doi:10.1007/s00184-008-0193-4

Number of Papers published in peer-reviewed journals: 11.00

(b) Papers published in non-peer-reviewed journals or in conference proceedings (N/A for none)

These are basically invited book chapters.

1. Wegman, Edward J. and Chow, Winston (2004) “Modeling continuous time series driven by fractional Gaussian noise,” in Time Series Analysis and Applications to Geophysical Systems, a book in the series : The IMA Volumes in Mathematics and its Applications , Vol. 139, New York: Springer-Verlag, (Brillinger, David R.; Robinson, Enders A.; Schoenberg, Frederic P., Eds.), 239-256

2. Solka, J.L., Adams, M.L., and Wegman, E.J. (2004) “Man vs. machine - A study of the ability of statistical methodologies to discern human generated ssh traffic from machine generated scp traffic,” in Statistical Methods in Computer Security, (W. Chen, ed.), Marcel-Dekker, New York, 169-181

3. Wegman, Edward J. (2005) “On some statistical methods for parallel computation,” Handbook of Parallel Computing and Statistics, (Erricos John, Ed.) 285-307

4. Marchette, David J., Wegman, Edward J. and Priebe, Carey E. (2005) “Fast algorithms for classification using class cover digraphs,” Handbook of Statistics: Data Mining and Data Visualization, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 331-358

5. Wegman, Edward J. and Solka, Jeffrey L.(2005) “Statistical data mining,” Handbook of Statistics: Data Mining and Data Visualization, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 1-46

6. Solka, Jeffrey L., Bryant, Avory C., and Wegman, Edward J. (2005) “Text data mining with minimal spanning trees,” Handbook of Statistics: Data Mining and Data Visualization, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 133-170

7. Moustafa, Rida E. A. and Wegman, Edward J. (2006) “Multivariate continuous data, generalizations of parallel coordinates,” Graphics of Large Datasets: Visualizing a Million, (Antony Unwin, Martin Theus, Heike Hofmann, eds.) 143-156

Number of Papers published in non peer-reviewed journals: 7.00

(c) Presentations

1.“Visual Data Mining of Streaming Data,” Federal Committee on Statistical Methodology (FCSM Statistical Policy Seminar: Achieving Statistical Quality in a Diverse and Changing Environment, Bethesda, MD, December, 2004

2. “Cybersecurity on the Internet: Where the Homeland is the World,” Public Lecture, International Conference on the Future of Statistical Theory, Practice and Education, Hyderabad, India, December 2004-January 2005

3. “Ecology of Alcohol and Alcoholism,” International Conference on the Future of Statistical Theory, Practice and Education, Hyderabad, India, December 2004-January 2005

4. “The Development and Implications of Computational Statistics for Social Science, Health and Other Applications,” Keynote Talk, Milestones in 21st Century Science, Buffalo, NY, March, 2005

5. “Strategies for Visual Data Mining,” Keynote Talk, SIAM Data Mining Conference 2005, Orange County, California, April, 2005

6. “40 Years of Statistics Research: A Personal Perspective,” 40 years of Statistical Computing and Beyond, Murray Hill, NJ, April, 2005

7. “Strategies for Visual Data Mining,” Symposium on the Interface, St. Louis, MO, June, 2005

8. “Ecology of Alcohol and Alcoholism,” with Y. Said, Symposium on the Interface, St. Louis, MO, June, 2005

9. “Visual Data Mining,” Introductory Overview Lecture, Joint Statistical Meetings, Minneapolis, MN, August, 2005

10. “Automated Metadata,” SAMSI Workshop on Homeland Defense and National Security, Research Triangle Park, NC, September, 2005

11. “Automated Metadata,” Army Conference on Applied Statistics, Monterey, CA, October, 2005

12. “Automated Metadata for Text Mining,” ASA/RAND Conference on Quantitative Methods & Statistical Applications in Defense, Santa Monica, CA, February, 2006

13. “Statistics, Data Mining, and Climate Change,” Keynote Talk, Second NASA Datamining Workshop: Issues and Applications in Earth Science, Pasadena, CA, May, 2006

14. “Statistics, Data Mining, and Climate Change,” Keynote Talk, Symposium on the Interface, Pasadena, CA, May, 2006

15. “Text Data Mining with Minimal Spanning Trees,” Summer Research Conference on Statistics, Kerrville, TX, June, 2006

16. Testimony to House Committee on Energy and Commerce, U.S. House of Representatives, Washington, D.C., July 20 and 27, 2006, http://republicans.energycommerce.house.gov/108/home/07142006_Wegman_fact_sheet.pdf
http://republicans.energycommerce.house.gov/108/News/07142006_1990.htm

17. “The Kyoto Accord, The 2001 IPCC Third Assessment Report and The Academic Papers Underpinning Them,” Joint Statistics Meeting, Seattle, WA, August, 2006

18. “Density Estimation from Streaming Data Using Wavelets,” COMPSTAT 2006, Rome, Italy, August, 2006

19. “Geospatial Distribution of Alcohol-Related Violence in Northern Virginia,” COMPSTAT 2006, Rome, Italy, August 2006

20. “On the Extraction of Endogenous Metadata for Text and Image Databases,” Keynote Talk, The KNEMO Workshop, Anacapri, Capri, Italy, September 2006

21. “Computational Statistics – Graphical and Analytic Methods for Streaming Data,” Short Course Lectures, Universita Napoli “Federico II”, Naples, Italy, September, 2006

22. “The Hockey Stick Controversy: Lessons for Statisticians,” Army Conference on Applied Statistics, Research Triangle Park, NC, October, 2006

23. “Visual Data Mining,” Public Lecture, Distinguished Visiting Professor at the American University of Cairo, Cairo, Egypt, March, 2007

24. “Visualization of Streaming Data,” Public Lecture, Distinguished Visiting Professor at the American University of Cairo, Cairo, Egypt, March, 2007

25. “Reanalysis of the Hockey Stick Paleoclimate Reconstruction,” Public Lecture, Distinguished Visiting Professor at the American University of Cairo, Cairo, Egypt, March, 2007

26. “Extraction of Endogenous Metadata,” Keynote Talk, Sixth Conference on Statistics in the Social Sciences and Humanities, Cairo, Egypt, March, 2007

27. “Assessing Interventions Related to the Negative Effects of Ethanol on HIV/AIDS Spread,” 39th Symposium on the Interface of Computing and Statistics, Philadelphia, PA, May, 2007

28. “A Bipartite Graph Model of the Interaction between Alcohol Users and Institutions,” Research Society on Alcoholism Annual Meeting, Chicago, IL, July, 2007

29. “Assessing Interventions Related to HIV Incidents Under the Influence of Ethanol,” Joint Statistical Meetings, Salt Lake City, UT, August, 2007

30. “Text Mining for Fun and Profit,” 3rd International Symposium on Business and Industrial Statistics, Ponta Delgada, Azores, Portugal, August, 2007

31. “20 Questions a Statistician Should Ask about Climate Change,” ASA Workshop on Climate Change, NCAR, Boulder, CO, October, 2007

32. “Methods for Visualizing High Dimensional Data,” Contemporary Frontiers in High Dimensional Statistical Analysis, Cambridge, UK, January, 2008

33. “Text Mining, Social Networks, and High Dimensional Analysis,” Izzet Sahin Memorial Lecture, University of Wisconsin, Milwaukee, WI, April, 2008

34. “Text Mining and Social Networks: Some Unexpected Connections,” Keynote Address, International Conference on Multivariate Statistical Modeling and High Dimensional Data Mining, Kayseri, Turkey, June, 2008.

35. “Approaches to Text Mining that Preserve Semantic Content,” Yasmin H. Said, Lecturer, International Conference on Multivariate Statistical Modeling and High Dimensional Data Mining, Kayseri, Turkey, June, 2008.

36. “Mixture Models for Document Clustering,” Joint Statistical Meetings, Denver, CO, August, 2008

Number of Presentations: 36.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

1. Faxon, Don, King, R. Duane, Rigsby, John T., Bernard, Steve, and Wegman, Edward J. (2004) "Data cleansing and preparation at the gates: A data-streaming perspective." With, In 2004 Proceedings of the American Statistical Association

2. Alnoshan, Abdullah, Rotenstreich, Shmuel, Wegman, Edward, Said, Yasmin and Rajput, Adil (2007) "Microeconomic approach to resource allocation in P2P grids," Proceedings of the Joint Statistical Meetings, 1975-1980

3. Sharabati, Walid K., Wegman, Edward J. and Said, Yasmin H. (2007) "A model of preferential attachments for emerging scientific subfields," Proceedings of the Joint Statistical Meetings, 2048-2055

4. Said, Yasmin H. and Wegman, Edward J. (2007) "Restrictions of trans fatty acids: Health benefits and economic impact in the Washington, DC Metro Area," Proceedings of the Joint Statistical Meetings, 1523-1527

5. Mburu, Peter K., Said, Yasmin H. and Wegman, Edward J. (2007) "Temporal statistics for consequences of alcohol use," Proceedings of the Joint Statistical Meetings, 2005-2009

6. Lin, Chien-Chih, Noh, Eun Young, Yan, Younggping, and Wegman, Edward J. (2008) "User profiling in window title and process table," Computing Science and Statistics, 36, 530-546

7. Martinez, Wendy L., Martinez, Angel R. and Wegman, Edward J. (2008) "Classification and clustering using weighted text proximity matrices," Computing Science and Statistics, 36, 600-611

8. Solka, Jeffrey L., Bryant, Avory C. and Wegman, Edward J. (2008) "Identifying cross corpora document associations via minimal spanning trees," Computing Science and Statistics, 36, 952-961

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts): 8

Peer-Reviewed Conference Proceeding publications (other than abstracts):

1. Kafadar, Karen and Wegman, Edward J. (2004) "Graphical displays of Internet traffic data," COMPSTAT 2004, (Antoch, J., ed.), Berlin: Physica-Verlag, 287-302

2. Priebe, C.E., Marchette, D.J., Park, Y., Wegman, E.J., Solka, J.L., Socolinsky, D.A., Karakos, D., Church, K.W., Guglielmi, R., Coifman, R.R., Lin, D., Healey, D.M., Jacob, M.Q., and Tsao, A. (2004) "Iterative denoising for cross-corpus discovery," COMPSTAT 2004, (Antoch, J., ed.), Berlin: Physica-Verlag, 381-392

3. Martinez, A.R., Wegman, E.J. and Martinez, W.L. (2004) "Using weights with a text proximity matrix," COMPSTAT 2004, (Antoch, J., ed.), Berlin: Physica-Verlag, 327-338

4. Alotaiby, Fahad T., Chen, Jim X., Wegman, Edward J., Wechsler, Harry, and Sprague, Debra (2004) "Teacher-driven, web-based learning system," in Proceedings of the 5th Conference on Information Technology Education, ACM SIGITE, 284

5. Alotaiby, Fahad T., Chen, Jim X., Wechsler, Harry, Wegman, Edward J., and Sprague, Debra (2005) "Adaptive web-based learning system," in the Proceedings of the12th Annual IEEE International Conference and Workshop on the Engineering of Computer-Based Systems, 423-430

6. Said, Yasmin H and Wegman, Edward J. (2006) "Geospatial distribution of alcohol-related violence in Northern Virginia," in COMPSTAT 2006, (Alfredo Rizzi and Maurizio Vichi, eds.), 197-208

7. Wegman, Edward J. and Caudle, Kyle A. (2006) "Density estimation from streaming data using wavelets," in COMPSTAT 2006, (Alfredo Rizzi and Maurizio Vichi, eds.), 231-244

8. Said, Yasmin H., Wegman, Edward J., Sharabati, Walid K. and Rigsby, John T. (2007) "Implications of co-author networks on peer review," in Classification and Data Analysis, Macerata, Italy: EUM-Edizoni Università di Macerata, 245-248

9. Wegman, Edward J. and Said, Yasmin H. (2008) "A directed graph model of ecological alcohol systems incorporating spatiotemporal effects," COMPSTAT 2008, (Paula Brito, ed.), 179-190

10. Wieczorek, William F., Said, Yasmin H. and Wegman, Edward J. (2008) "Spatial and computational models of alcohol use and problems," COMPSTAT 2008, (Paula Brito, ed.), 191-202

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts): 10

(d) Manuscripts

1. Said, Yasmin H., Wegman, Edward J. and Sharabati, Walid K. (2008) "Author-coauthor spcial networks and emerging scientific subfields," to appear Data Analysis and Classification: From the Exploratory to the Confirmatory Approach, (Carlo Lauro, Francesco Palumbo, Michael Greenacre eds.) Berlin: Springer-Verlag

2. Wegman, Edward J. and Said, Yasmin H. (2008) "Text mining with application to fraud discovery," submitted to Applied Stochastic Models in Business and Industry

3. Said, Yasmin H. and Wegman, Edward J. (2008) "Agent-based simulation of the alcohol ecological system," submitted Journal of the American Statistical Association

Number of Manuscripts: 3.00

Number of Inventions:

Graduate Students

NAME	PERCENT_SUPPORTED
Yasmin Said	0.25
Eun Noh	0.50
FTE Equivalent:	**0.75**
Total Number:	**2**

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Yasmin Said	0.25
FTE Equivalent:	**0.25**
Total Number:	**1**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Edward Wegman	0.33	No
FTE Equivalent:	**0.33**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	0.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):	0.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME
John Rigsby
Total Number: 1

Names of personnel receiving PHDs

NAME
Yasmin Said
Kyle Caudle
Eun Noh
Faleh Alshameri
Fahad Alotaiby
Homayoun Sharafi
Elizabeth Hohman
Total Number: 7

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

5 **A Graph-Theoretic Policy Decision Tool for Exploring Interventions**

Patent Filed in US? (5d-1) Y

Patent Filed in Foreign Countries? (5d-2) N

Was the assignment forwarded to the contracting officer? (5e) N

Foreign Countries of application (5g-2):

5a: Edward J. Wegman

5f-1a: George Mason University

5f-c: 4400 University Drive

Fairfax VA 22030

5a: Yasmin H. Said

5f-1a: George Mason University

5f-c: 4400 University Drive

Fairfax VA 22030

5 **Automated Generation of Metadata for Mining Text and Image Data**

Patent Filed in US? (5d-1) Y

Patent Filed in Foreign Countries? (5d-2) N

Was the assignment forwarded to the contracting officer? (5e) N

Foreign Countries of application (5g-2):

5a: Faleh Jassem Al-Shameri

5f-1a: Howard University

5f-c: 2400 6th St. NW

Washington DC 20059

5a: Edward J. Wegman

5f-1a: George Mason University

5f-c: 4400 University Drive

Fairfax VA 22039

Analytic and Graphical Methods for Streaming Data with Applications to Netcentric Warfare - Final Report

Background

Netcentric warfare concepts evolved from the 1991 Gulf War experience and have been defined by several sources. The Committee on Network-Centric Naval Forces of the Naval Studies Board of the National Research Council (2000) defined Network Centric Operations as follows:

> "Network-centric operations (NCO) [are] military operations that exploit state-of-the-art information and networking technology to integrate widely dispersed human decision makers, situational and targeting sensors, and forces and weapons into a highly adaptive, comprehensive system to achieve unprecedented mission effectiveness."

They go on to observe that:

> "In one way or another all military operations will be joint. That is, systems and forces from all the services and from National agencies will contribute to the U. S. Armed Forces' operations in ways that vary with the circumstances."

The reliance for total mission effectiveness on information and networking technology brings concomitant vulnerabilities. Modern information and networking technology could be destroyed relatively easily by an electro-magnetic pulse. More likely however, is the interception of networking technology by clever hackers, even on secure communication networks. Much of the communication technology especially from sensor platforms is radio based and subject to spoofing and other methods of distorting the overall situational awareness.

Data mining, when compared with the classical statistical analysis paradigm, shows a substantial change of perspective. Instead of relatively small, low-dimensional, homogeneous data sets derived from a carefully designed sampling procedure, awareness of the issues in data mining has led many researchers to consider massive, high-dimensional datasets that may not satisfy traditional homogeneity assumptions. In addition, data sets used in a data mining context often have been collected for administrative or other purposes having little to do with the purposes for which data mining techniques are applied. Nonetheless, even among those aware of the issues of computational complexity and massive data sets, the typical mindset is that we have a dataset of fixed size n, however large n might be. However, we believe there is a new data collection paradigm looming on the horizon, to wit, data are streaming, coming available continuously, and realistically not all of it can be stored. It is clear that the netcentric operations idea will generate a streaming of the type suggested by what we identify as the new paradigm. Moreover as new data becomes available the value of older data diminishes. Almost a given with streaming data is that data are not time homogeneous. Indeed this is a strong contrast with the conventional perspective on homogeneously sampled data.

In addition to data generated by networks of computers, examples of this class of data abound. Point of purchase data, telephone traffic data, and credit card use data are all examples. The data on which we will focus in this proposal is Internet traffic data. Techniques for the use and analysis of such data must of necessity be somewhat different from fixed sample size data. Because the data cannot be stored conveniently, recursive algorithms that update the desired statistic using an incoming piece of data and then discard that piece of data are appropriate. Data visualization methods have more recently emphasized dynamic graphics, i.e. animation of the graphic using rotation, grand tours, mapping of variables into time, and so on, but always with an eye on a fixed dataeset size. We propose what we like to call *evolutionary graphics*, i.e. graphics which evolve as a function of new data being added. The combination of recursive algorithms and evolutionary graphics will provided a fundamental approach for analyzing streaming data.

As a prototype for developing tools for streaming data, we have launched on a data collection effort with the agreement of the George Mason University's CIO to capture all header information for all Internet traffic in and out of the University. This comprises primarily TCP, UDP and ICMP packets. We have installed sniffer and analysis machines and are capable of locally recording up to a terabyte of traffic data. Preliminary experiments within our small statistics subnet indicate traffic of 65,000 to 150,000 packets per hour. We are currently collecting about 1.2 to 3.0 gigabytes of header information per hour in the larger University context. Ultimately, we are interested in real-time detection of intrusion attacks so that analysis methods for steaming data are necessary. In the next sections, we will describe some background on TCP/IP traffic, indicate some recursive methods capable of handling streaming data and give some suggestions for analytic algorithms and visualization procedures we hope to develop under this proposal.

Statement of Problem

The effort of this project is directed toward developing both graphical and analytic methods for streaming data. While traditional data analysis paradigms assume a fixed sample size and proceed to develop analysis tools based on assumptions of homogeneity in the mechanism generating the data, more and more the ease of electronic acquisition of data implies that data are streaming and that data is likely to be non-homogeneous. Examples include Internet traffic, battlefield communications, financial transactions, news stories, and satellite remote sensing to name a few. The data may be categorical, numerical, text documents or imagery as these examples imply. A strong motivation for investigating analysis of streaming data is the increasing digitization of the battlefield and the migration to net-centric warfare. The possibility of malicious interception of packet data and of spoofing legitimate communications suggests that quick response analysis and visualization tools in order to detect atypical behavior is exceptionally useful in preventing loss of information in a warfare setting.

One major implication of streaming data is that ultimately there is not sufficient storage to store all of the data, so that data must ultimately be discarded. More importantly, there

is a further implication that older data is of less value for making current inferences. Our general approach has been to develop recursive algorithms that are able to process data in one pass. Approaches for implementing recursive algorithms include: evolutionary graphics, i.e. graphics, which evolve as a function of new data being added, implementation of a quantizing algorithm for truly massive streaming data, recursive kernel density estimators in the critical multidimensional case, adaptive mixtures density estimation algorithms using orthonormal bases such as wavelets to reduce or eliminate spurious terms, and classical exponential smoother with adaptive time scaling and their potential use as multiscale data analysis. Our approach is distinctly different from classical time series analysis methodology in that generally classical time series analysis assumes stationarity or at least systematic nonstationarity. Generally our focus has been on digital TCP/IP traffic headers as a proxy for all packet-based communications protocols.

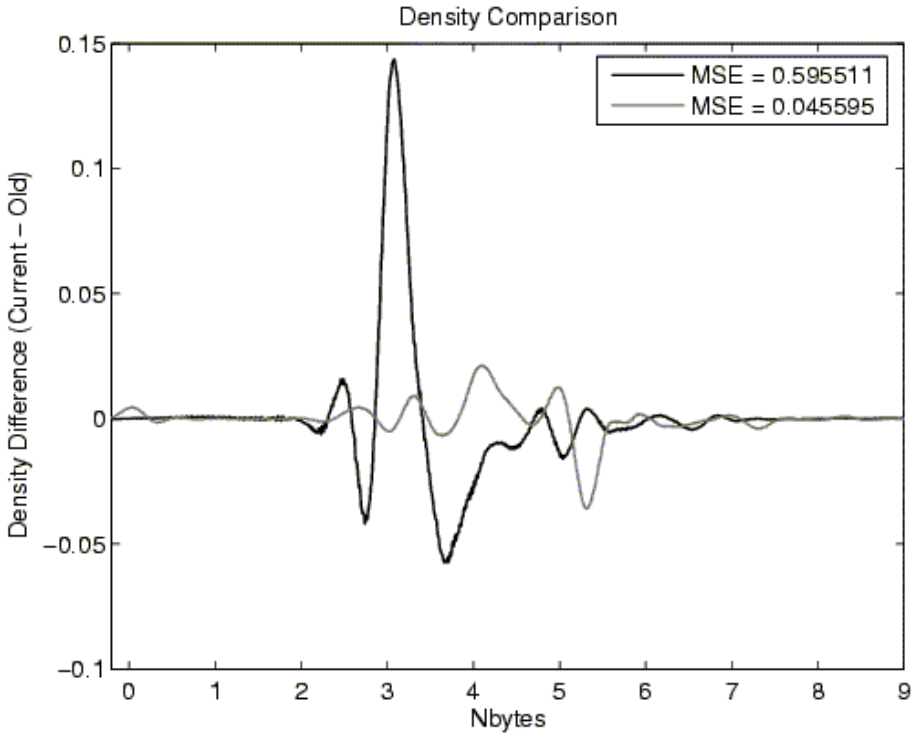

Figure 1: Recursively formulated difference density estimates.

Significance: The recursive methods for both univariate and multivariate density estimation have proven their ability to detect subtle changes in patterns of IP traffic. We are able to track changes within 1000 observations, which is a matter of microseconds in the Internet traffic stream. Figure 1 illustrates this point. The curve labeled $MSE = 0.045595$ represents the difference between two density estimates based on steaming data, which are calculated using two 1000 observation samples when the samples have no intrusion attempt. The curve labeled $MSE = 0.59551$ represents the

difference between two density estimates calculated using two 1000 observation samples when there was an intrusion attempt. The implication of this is that while packet traffic may change in a non-stationary fashion, the time scale of intruders in to the system is much shorter than the time scale for changes of the general system. The translation of these methods to operational systems would imply an ability to give alerts rapidly when patterns change, thus real-time or near real-time sensitivity to attempts to compromise the system. The evolutionary graphics have also shown great promise although the implementation suffers from the lack of continuously available data. One interesting aspect is that the evolutionary visualization of source IP suggests characteristic patterns of users of the system. In cases where there is a dedicated IP, such variations can be used to detect unauthorized users of a system, i.e. possible insider threat.

The concept of evolutionary graphics, which we also explored, takes the notion of dynamic graphics one step further by allowing a graphic to change real time with streaming data. Figure 2 illustrated below is a snapshot of a "waterfall diagram." In this diagram we are dynamicall recording source IP addresses so that we may monitor where IP traffic coming into our site is originating. This allows us to identify possible hostile attempts to intrude.

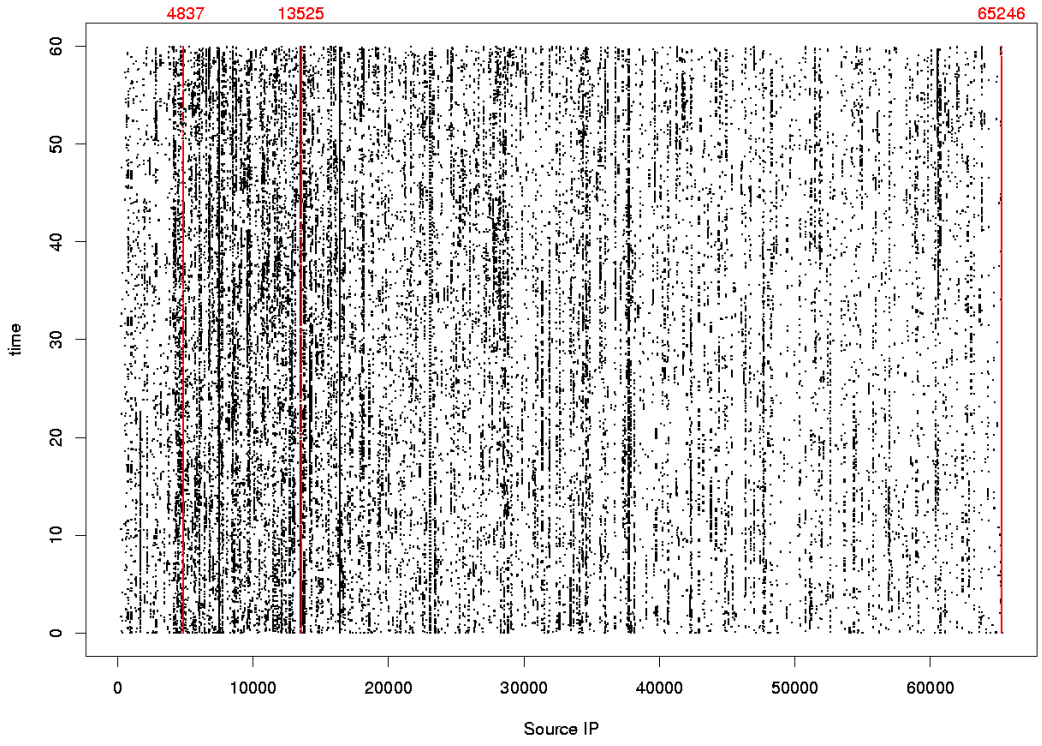

Figure 2: Source IP waterfall diagram.

In a similar way we can form a waterfall diagram with Source Port information instead of Source IP. The evolution of source port increments is characteristic of operating systems. The various slopes in Figure 3 are characterizing different systems.

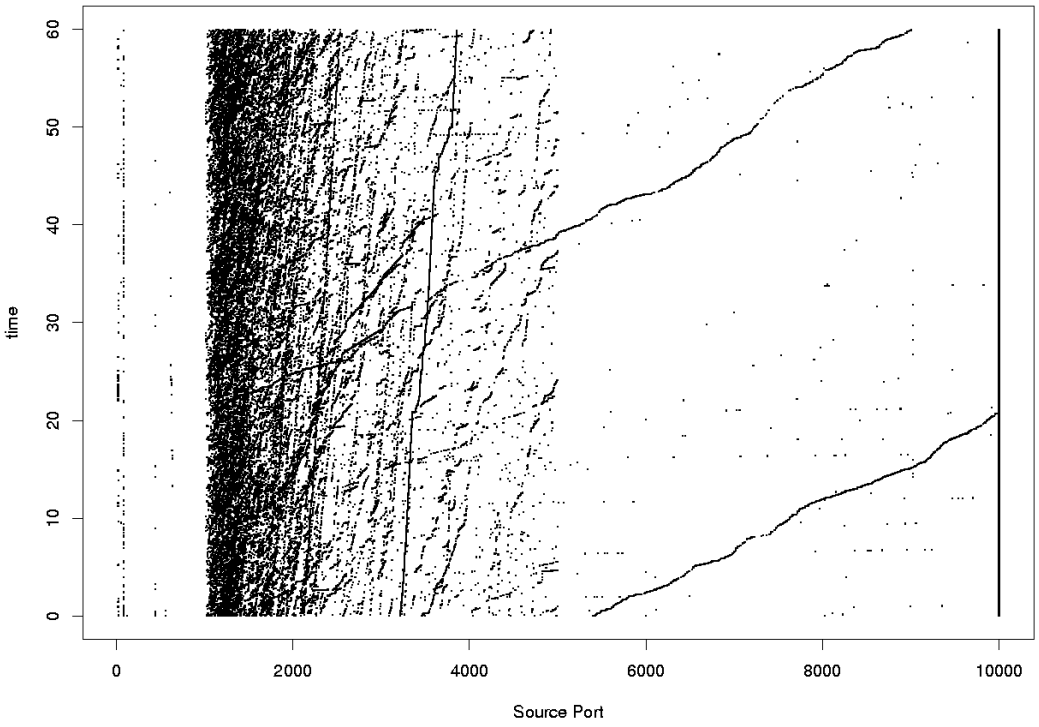

Figure 3: Source Port waterfall diagram.

Another evolutionary graphic that we have been developing, we called transient geographic mapping. The idea is that we geolocate the source of an IP address and flash it onto a map. The point of light is allowed to decay with a controllable decay rate. If we receive many packets from a given location we will see a sustained intensity. If we see much traffic from a hostile location, it suggests that further investigation is needed. Figure 4 is a screen shot from the application that was developed.

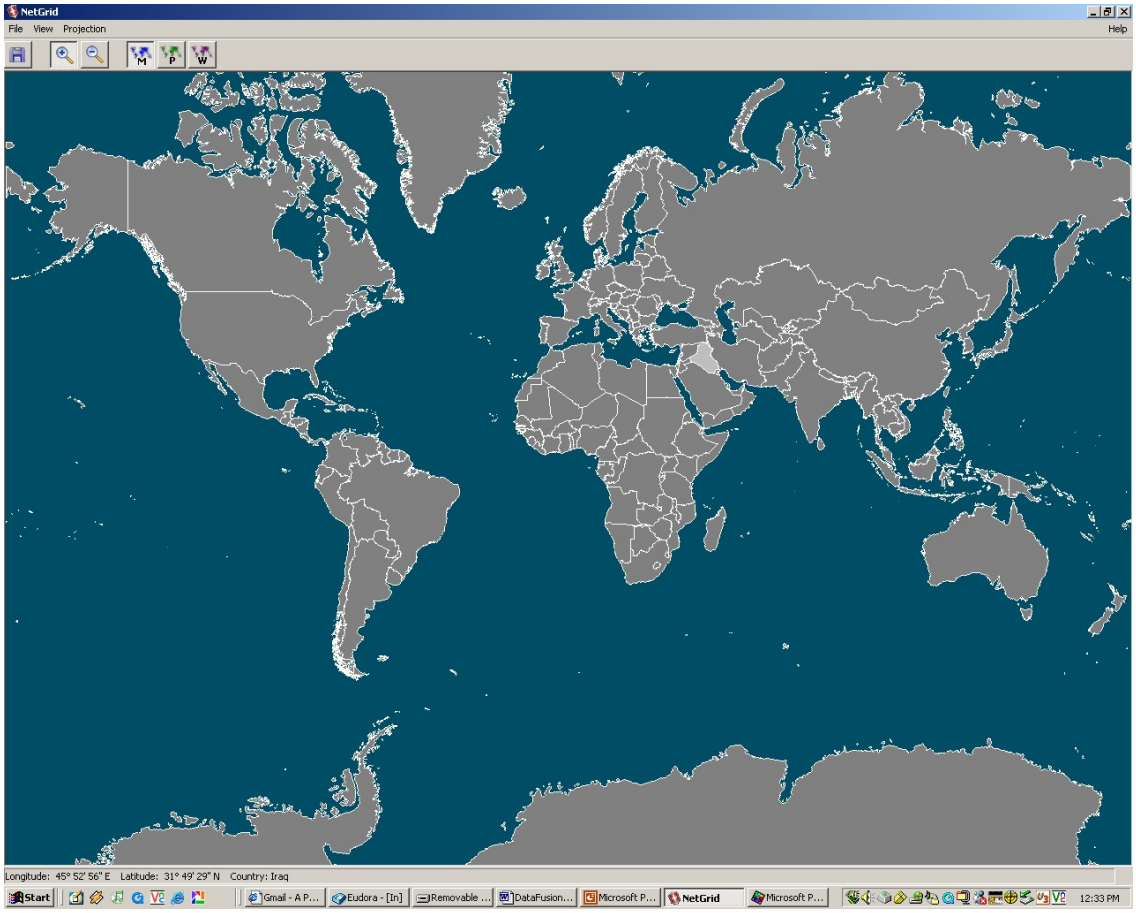

Figure 4: Transient Geographic Mapping screen shot.

Further Work: As the project has evolved, we began an increasing focus on the network aspects of the netcentric activities. This has involved computer as well as social networks. An interesting aspect of our work with social networks was to realize that the adjacency matrix for the so-called two-mode social network has exactly the same structure as the term-document matrix in computational linguistics and text data mining. Consequently, we have pursued some ideas in text mining and that had evolved into an interest in multilingual text mining. Figure 5 represents an architecture for a multilingual text mining system we have begun to develop.

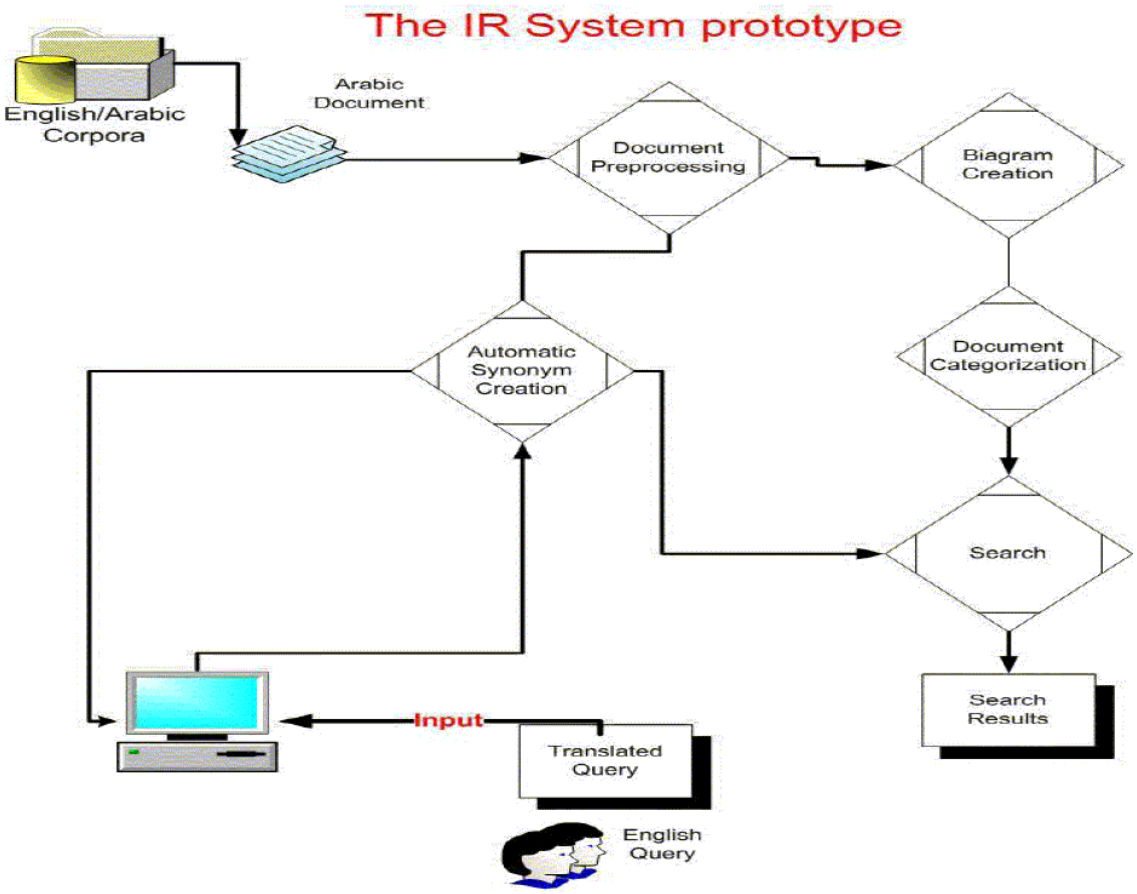

Figure 5: Architecture for a multi-lingual text mining system.

A complete list of talks and papers on these subjects is given below. Individual papers are available upon request.

Invited Addresses by Edward J. Wegman Acknowledging Support of ARO Contract W911NF-04-1-0447

- "Visual Data Mining of Streaming Data," Federal Committee on Statistical Methodology (FCSM Statistical Policy Seminar: Achieving Statistical Quality in a Diverse and Changing Environment, Bethesda, MD, December, 2004
- "Cybersecurity on the Internet: Where the Homeland is the World," Public Lecture, International Conference on the Future of Statistical Theory, Practice and Education, Hyderabad, India, December 2004-January 2005
- "Ecology of Alcohol and Alcoholism," International Conference on the Future of Statistical Theory, Practice and Education, Hyderabad, India, December 2004-January 2005
- "The Development and Implications of Computational Statistics for Social Science, Health and Other Applications," Keynote Talk, Milestones in 21st Century Science, Buffalo, NY, March, 2005
- "Strategies for Visual Data Mining," Keynote Talk, SIAM Data Mining Conference 2005, Orange County, California, April, 2005
- "40 Years of Statistics Research: A Personal Perspective," 40 years of Statistical Computing and Beyond, Murray Hill, NJ, April, 2005
- "Strategies for Visual Data Mining," Symposium on the Interface, St. Louis, MO, June, 2005
- "Ecology of Alcohol and Alcoholism," with Y. Said, Symposium on the Interface, St. Louis, MO, June, 2005
- "Visual Data Mining," Introductory Overview Lecture, Joint Statistical Meetings, Minneapolis, MN, August, 2005
- "Automated Metadata," SAMSI Workshop on Homeland Defense and National Security, Research Triangle Park, NC, September, 2005
- "Automated Metadata," Army Conference on Applied Statistics, Monterey, CA, October, 2005
- "Automated Metadata for Text Mining," ASA/RAND Conference on Quantitative Methods & Statistical Applications in Defense, Santa Monica, CA, February, 2006
- "Statistics, Data Mining, and Climate Change," Keynote Talk, Second NASA Datamining Workshop: Issues and Applications in Earth Science, Pasadena, CA, May, 2006
- "Statistics, Data Mining, and Climate Change," Keynote Talk, Symposium on the Interface, Pasadena, CA, May, 2006
- "Text Data Mining with Minimal Spanning Trees," Summer Research Conference on Statistics, Kerrville, TX, June, 2006
- Testimony to House Committee on Energy and Commerce, U.S. House of Representatives, Washington, D.C., July 20 and 27, 2006, http://republicans.energycommerce.house.gov/108/home/07142006_Wegman_fact_sheet.pdf http://republicans.energycommerce.house.gov/108/News/07142006_1990.htm
- "The Kyoto Accord, The 2001 IPCC Third Assessment Report and The Academic Papers Underpinning Them," Joint Statistics Meeting, Seattle, WA, August, 2006
- "Density Estimation from Streaming Data Using Wavelets," COMPSTAT 2006, Rome, Italy, August, 2006

- "Geospatial Distribution of Alcohol-Related Violence in Northern Virginia," COMPSTAT 2006, Rome, Italy, August 2006
- "On the Extraction of Endogenous Metadata for Text and Image Databases," Keynote Talk, The KNEMO Workshop, Anacapri, Capri, Italy, September 2006
- "Computational Statistics – Graphical and Analytic Methods for Streaming Data," Short Course Lectures, Universita Napoli "Federico II", Naples, Italy, September, 2006
- "The Hockey Stick Controversy: Lessons for Statisticians," Army Conference on Applied Statistics, Research Triangle Park, NC, October, 2006
- "Visual Data Mining," Public Lecture, Distinguished Visiting Professor at the American University of Cairo, Cairo, Egypt, March, 2007
- "Visualization of Streaming Data," Public Lecture, Distinguished Visiting Professor at the American University of Cairo, Cairo, Egypt, March, 2007
- "Reanalysis of the Hockey Stick Paleoclimate Reconstruction," Public Lecture, Distinguished Visiting Professor at the American University of Cairo, Cairo, Egypt, March, 2007
- "Extraction of Endogenous Metadata," Keynote Talk, Sixth Conference on Statistics in the Social Sciences and Humanities, Cairo, Egypt, March, 2007
- "Assessing Interventions Related to the Negative Effects of Ethanol on HIV/AIDS Spread," 39th Symposium on the Interface of Computing and Statistics, Philadelphia, PA, May, 2007
- "A Bipartite Graph Model of the Interaction between Alcohol Users and Institutions," Research Society on Alcoholism Annual Meeting, Chicago, IL, July, 2007
- "Assessing Interventions Related to HIV Incidents Under the Influence of Ethanol," Joint Statistical Meetings, Salt Lake City, UT, August, 2007
- "Text Mining for Fun and Profit," 3rd International Symposium on Business and Industrial Statistics, Ponta Delgada, Azores, Portugal, August, 2007
- "20 Questions a Statistician Should Ask about Climate Change," ASA Workshop on Climate Change, NCAR, Boulder, CO, October, 2007
- "Methods for Visualizing High Dimensional Data," Contemporary Frontiers in High Dimensional Statistical Analysis, Cambridge, UK, January, 2008
- "Text Mining, Social Networks, and High Dimensional Analysis," Izzet Sahin Memorial Lecture, University of Wisconsin, Milwaukee, WI, April, 2008
- "Text Mining and Social Networks: Some Unexpected Connections," Keynote Address, International Conference on Multivariate Statistical Modeling and High Dimensional Data Mining, Kayseri, Turkey, June, 2008.
- "Approaches to Text Mining that Preserve Semantic Content," Yasmin H. Said, Lecturer, International Conference on Multivariate Statistical Modeling and High Dimensional Data Mining, Kayseri, Turkey, June, 2008.
- "Mixture Models for Document Clustering," Joint Statistical Meetings, Denver, CO, August, 2008

Papers Published by Edward J. Wegman Acknowledging ARO Support under Contract W911NF-04-1-0447 (or its predecessor contract DAAG55-98-1-0404, but not report in the predecessor final report)

- Solka, Jeffrey L., Wegman, Edward J., and Marchette, David J. (2004) "Data mining strategies for detection of chemical warfare agents," *Statistical Data Mining and Knowledge Discovery*, 71-92
- Marchette, David J. and Wegman, Edward J. (2004) "Statistical analysis of network data for cybersecurity," *Chance*, 17(1), 8-18
- Wegman, Edward J. and Chow, Winston (2004) "Modeling continuous time series driven by fractional Gaussian noise," in *Time Series Analysis and Applications to Geophysical Systems*, a book in the series : The IMA Volumes in Mathematics and its Applications , Vol. 139, New York: Springer-Verlag, (Brillinger, David R.; Robinson, Enders A.; Schoenberg, Frederic P., Eds.), 239-256
- Johannsen, D.A., Wegman, E.J., Solka, J.L. and Priebe, C.E. (2004) "Simultaneous selection of features and metric for optimal nearest neighbor classification," *Communications in Statistics: Theory and Methods*, 2137-2158
- Kafadar, Karen and Wegman, Edward J. (2004) "Graphical displays of Internet traffic data," *COMPSTAT 2004*, (Antoch, J., ed.), Berlin: Physica-Verlag, 287-302
- Priebe, C.E., Marchette, D.J., Park, Y., Wegman, E.J., Solka, J.L., Socolinsky, D.A., Karakos, D., Church, K.W., Guglielmi, R., Coifman, R.R., Lin, D., Healey, D.M., Jacob, M.Q., and Tsao, A. (2004) "Iterative denoising for cross-corpus discovery," *COMPSTAT 2004*, (Antoch, J., ed.), Berlin: Physica-Verlag, 381-392
- Martinez, A.R., Wegman, E.J. and Martinez, W.L. (2004) "Using weights with a text proximity matrix," *COMPSTAT 2004*, (Antoch, J., ed.), Berlin: Physica-Verlag, 327-338
- Faxon, Don, King, R. Duane, Rigsby, John T., Bernard, Steve, and Wegman, Edward J. (2004) "Data cleansing and preparation at the gates: A data-streaming perspective." With, In *2004 Proceedings of the American Statistical Association*
- Solka, J.L., Adams, M.L., and Wegman, E.J. (2004) "Man vs. machine - A study of the ability of statistical methodologies to discern human generated ssh traffic from machine generated scp traffic," in *Statistical Methods in Computer Security*, (W. Chen, ed.), Marcel-Dekker, New York, 169-181
- Alotaiby, Fahad T., Chen, Jim X., Wegman, Edward J., Wechsler, Harry, and Sprague, Debra (2004) "Teacher-driven, web-based learning system," in *Proceedings of the 5th Conference on Information Technology Education*, ACM SIGITE, 284
- Wegman, Edward J. (2005) "On some statistical methods for parallel computation," *Handbook of Parallel Computing and Statistics*, (Erricos John, Ed.) 285-307
- Marchette, David J., Wegman, Edward J. and Priebe, Carey E. (2005) "Fast algorithms for classification using class cover digraphs," *Handbook of Statistics: Data Mining and Data Visualization*, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 331-358
- Wegman, Edward J. and Solka, Jeffrey L.(2005) "Statistical data mining," *Handbook of Statistics: Data Mining and Data Visualization*, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 1-46
- Solka, Jeffrey L., Bryant, Avory C., and Wegman, Edward J. (2005) "Text data mining with minimal spanning trees," *Handbook of Statistics: Data Mining and Data Visualization*, (Rao, C. R., Wegman, E. J. and Solka, J. L., eds.), 133-170

- Alotaiby, Fahad T., Chen, Jim X., Wechsler, Harry, Wegman, Edward J., and Sprague, Debra (2005) "Adaptive web-based learning system," in the Proceedings of the *12th Annual IEEE International Conference and Workshop on the Engineering of Computer-Based Systems*, 423-430
- Moustafa, Rida E. A. and Wegman, Edward J. (2006) "Multivariate continuous data, generalizations of parallel coordinates," *Graphics of Large Datasets: Visualizing a Million*, (Antony Unwin, Martin Theus, Heike Hofmann, eds.) 143-156
- Kafadar, Karen and Wegman, Edward J. (2006) "Visualizing 'typical' and 'exotic' Internet traffic data," *Computational Statistics and Data Analysis*, 50(12), 3721-3743
- Said, Yasmin H and Wegman, Edward J. (2006) "Geospatial distribution of alcohol-related violence in Northern Virginia," in *COMPSTAT 2006*, (Alfredo Rizzi and Maurizio Vichi, eds.), 197-208
- Wegman, Edward J. and Caudle, Kyle A. (2006) "Density estimation from streaming data using wavelets," in *COMPSTAT 2006*, (Alfredo Rizzi and Maurizio Vichi, eds.), 231-244
- Dorfman, Alan H., Lent, Janice, Leaver, Sylvia G. and Wegman, Edward J. (2006) "On sample survey designs for consumer price indexes," *Survey Methodology*, 32(2), 197-216
- Said, Yasmin H., Wegman, Edward J., Sharabati, Walid K. and Rigsby, John T. (2007) "Implications of co-author networks on peer review," in *Classification and Data Analysis*, Macerata, Italy: EUM-Edizoni Università di Macerata, 245-248
- Said, Yasmin H., Wegman, Edward J., Sharabati, Walid K. and Rigsby, John T. (2008) "Style of author-coauthor social networks," *Computational Statistics and Data Analysis*, 52, 2177-2184, 2008; doi:10.1016/j.csda.2007.07.021, 2007
- Said, Yasmin H. and Wegman, Edward J. (2007) "Quantitative assessments of alcohol-related outcomes," *Chance*, 20(3), 17-25
- Wegman, Edward J. and Martinez, Wendy L. (2007) "A conversation with Dorothy Gilford," *Statistical Science*, 22(2), 291-300
- Alnoshan, Abdullah, Rotenstreich, Shmuel, Wegman, Edward, Said, Yasmin and Rajput, Adil (2007) "Microeconomic approach to resource allocation in P2P grids," *Proceedings of the Joint Statistical Meetings*, 1975-1980
- Sharabati, Walid K., Wegman, Edward J. and Said, Yasmin H. (2007) "A model of preferential attachments for emerging scientific subfields," *Proceedings of the Joint Statistical Meetings*, 2048-2055
- Said, Yasmin H. and Wegman, Edward J. (2007) "Restrictions of trans fatty acids: Health benefits and economic impact in the Washington, DC Metro Area," *Proceedings of the Joint Statistical Meetings*, 1523-1527
- Mburu, Peter K., Said, Yasmin H. and Wegman, Edward J. (2007) "Temporal statistics for consequences of alcohol use," *Proceedings of the Joint Statistical Meetings*, 2005-2009
- Said, Yasmin H. and Wegman, Edward J. (2008) "Using administrative data to estimate cyclic effects of alcohol usage (refereed abstract)," *Alcoholism: Clinical and Experimental Research*, 32(6) Supplement, 139A
- Wegman, Edward J. and Said, Yasmin H. (2008) "Modeling spatiotemporal effects for acute outcomes in an alcohol system (refereed abstract)," *Alcoholism: Clinical and Experimental Research*, 32(6) Supplement, 140A, 2008
- Lin, Chien-Chih, Noh, Eun Young, Yan, Younggping, and Wegman, Edward J. (2008) "User profiling in window title and process table," *Computing Science and Statistics*, 36, 530-546

- Martinez, Wendy L., Martinez, Angel R. and Wegman, Edward J. (2008) “Classification and clustering using weighted text proximity matrices,” *Computing Science and Statistics*, 36, 600-611
- Solka, Jeffrey L., Bryant, Avory C. and Wegman, Edward J. (2008) “Identifying cross corpora document associations via minimal spanning trees,” *Computing Science and Statistics*, 36, 952-961
- Reyen, Salem S., Miller, John J. and Wegman, Edward J. (2008) “Separating a mixture of two normals with proportional covariances,” *Metrika,* doi:10.1007/s00184-008-0193-4
- Wegman, Edward J. and Said, Yasmin H. (2008) “A directed graph model of ecological alcohol systems incorporating spatiotemporal effects,” *COMPSTAT 2008*, (Paula Brito, ed.), 179-190
- Wieczorek, William F., Said, Yasmin H. and Wegman, Edward J. (2008) “Spatial and computational models of alcohol use and problems,” *COMPSTAT 2008*, (Paula Brito, ed.), 191-202
- Said, Yasmin H., Wegman, Edward J. and Sharabati, Walid K. (2008) “Author-coauthor spcial networks and emerging scientific subfields,” to appear *Data Analysis and Classification: From the Exploratory to the Confirmatory Approach*, (Carlo Lauro, Francesco Palumbo, Michael Greenacre eds.) Berlin: Springer-Verlag
- Wegman, Edward J. and Said, Yasmin H. (2008) “Text mining with application to fraud discovery,” submitted to *Applied Stochastic Models in Business and Industry*
- Said, Yasmin H. and Wegman, Edward J. (2008) “Agent-based simulation of the alcohol ecological system,” submitted *Journal of the American Statistical Association*

Dissertation and Theses Directed by Edward J. Wegman while under Contract W911NF-04-1-0447

- Yasmin H. Said, *Agent Based Simulation of Ecological Alcohol Systems* (PhD) – Yasmin was supported as a postdoc partly by Contract W911NF-04-1-0447. She has worked on contract W911NF-07-1-0059 with the Army Research Laboratory in our technology transfer effort. She currently holds an F32 award from NIH.
- Kyle Allman Caudle, *Non-Parametric Density Estimation ff Streaming Data Using Orthogonal Series* (PhD) – Kyle is a Lt. Commander in the Navy currently deployed to Bahrain. He normally a faculty member at the United States Naval Academy.
- Eun Young Noh, *Multivariate Recursive Kernel Density Estimator with Adjustable Discounting Of Old Data* (PhD) – Eun Noh was supported by Contract W911NF-04-1-0447.
- Faleh Jassem Alshameri*, Automated Metadata for Mining Image and Text Data* (PhD) – Faleh has a faculty appointment at Howard University.
- Fahad Alotaiby, *A Component-based Functional Model for E-Learning Systems* (PhD) – Fahad is Chair of the Computer Science Department in Imam University, Riyadh, Saudi Arabia.
- Homayoun Sharafi, *Barriers to Teaching Computing Courses from a Distance at Community Colleges* (D Arts) – Homayoun’s degree focuses on community college teaching, he currently teaches at Northern Virginia Community College.
- Elizabeth Leeds Hohman, *A Dynamic Graph Model for Representing Streaming Text Documents* (PhD) – Elizabeth is employed as a research scientist at the Naval Surface Warfare Center, Dahlgren Division.

- John Thomas Rigsby, *Block Models and Allegiance* (MS) – John is employed as a research scientist at the Naval Surface Warfare Center, Dahlgren Division.