

SANDIA REPORT

SAND2010-8715

Unlimited Release

Printed December 2010

Network Discovery, Characterization, and Prediction: A Grand Challenge LDRD Final Report

Philip Kegelmeyer, *PI and Editor*

David Rogers, *Deputy Program Manager*

Rick Contreras, *Systems Research Liaison*

Mark Huey, *Business Intelligence*

Bill Cook, *Program Manager*

Amy Bowen, *Administrator*

Curtis Johnson, *Information Systems Analysis Liaison*

Bruce Hendrickson, *PI (FY08)*

Brett Bader, *Discovery Lead*

Timothy Shead, *Titan Lead*

Sue Medeiros, *Data Lead*

Ron Oldfield, *Data Architectures Lead*

Scott Mitchell, *Forecasting Lead*

Brian Wylie, *Prototypes Lead*

Laura McNamara, *Human Factors Lead*

Richard Murphy, *Architecture Processor Lead*

With:

G. Ronald Anderson, Alisa Bandlow, Travis Bauer, Janine Bennett, Jonathan Berry, Peter Chew, Richard Colbaugh, Kerstan Cole, Warren Davis IV, Courtney Dornburg, Daniel Dunlavy, Nathan Fabian, Alla Fishman, Mark Foehse, Charles Gieseler, Kristin Glass, John Greenfield, Eric Goodman, John Harger, Mark Hollingsworth, Tameka Huff, Bryan Ingram, Robert Kerr, Tamara Kolda, Randall Laviolette, Vitus Leung, Laura Matzen, Judy Spomer, Jonathan McClain, William McLendon, III, Dan Nordman, Thomas Otahal, Philippe Pebay, Cynthia Phillips, Ali Pinar, David Robinson, Matthew Rocklin, Mark Sears, Jason Shepherd, Austin Silva, Eric Stanton, Susan Stevens-Adams, Laura Swiler, William Stubblefield, David Thompson, Anne Tomasi, Timothy Trucano, Joel Vaughan, Andrew Wilson, Alyson Wilson

Prepared by

Sandia National Laboratories

Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.



Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: <http://www.osti.gov/bridge>

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>



Network Discovery, Characterization, and Prediction: A Grand Challenge LDRD Final Report

Philip Kegelmeyer (PI and Editor)
Computer Sciences and Information Systems Department
Sandia National Laboratories
P.O. Box 969
Livermore, CA, 94551
wpk@sandia.gov

Abstract

This report is the final summation of Sandia's Grand Challenge LDRD project #119351, "Network Discovery, Characterization and Prediction" (the "NGC") which ran from FY08 to FY10. The aim of the NGC, in a nutshell, was to research, develop, and evaluate relevant analysis capabilities that address adversarial networks. Unlike some Grand Challenge efforts, that ambition created *cultural* subgoals, as well as technical and programmatic ones, as the insistence on "relevancy" required that the Sandia informatics research communities and the analyst user communities come to appreciate each others' needs and capabilities in a very deep and concrete way.

The NGC generated a number of technical, programmatic, and cultural advances, detailed in this report. There were new algorithmic insights and research that resulted in fifty-three refereed publications and presentations; this report concludes with an abstract-annotated bibliography pointing to them all. The NGC generated three substantial prototypes that not only achieved their intended goals of testing our algorithmic integration, but which also served as vehicles for customer education and program development. The NGC, as intended, has catalyzed future work in this domain; by the end it had already brought in, in new funding, as much funding as had been invested in it. Finally, the NGC knit together previously disparate research staff and user expertise in a fashion that not only addressed our immediate research goals, but which promises to have created an enduring cultural legacy of mutual understanding, in service of Sandia's national security responsibilities in cybersecurity and counter proliferation.

Contents

Summary	6
1 Overview	9
1.1 The Problem	9
1.2 The Goals	9
1.3 The Project Principles	10
1.4 The Project Structure	12
2 Activities	15
2.1 Advances in Network Analysis	15
2.2 Engaging with Analysts Via Human Factors	18
2.3 Processor and Data Architecture Insights	26
2.4 Titan as the Enduring Computational Framework	28
2.5 The Three Prototypes	31
2.6 Acquiring and Wrestling with Data	37
2.7 Programmatic and Cultural Outcomes	39
3 Conclusion	42
3.1 Continuing Challenges and Opportunities	42
3.2 An Outside Perspective	43

Appendix

Annotated Bibliography	44
-------------------------------------	-----------

Figures

1 Dendrogram depictions of node partitioning	15
2 Tensor analysis of the Enron mail corpus	16
3 Correct detection of anomalous network events	17
4 Predictions from blog traffic	19
5 Example of a vertex edge graph	22
6 Example of a ring graph	23
7 The architectural solution for interactive multi-lingual document clustering	27
8 P1, illustrating almost all of its visualizations.	32
9 P2, illustrating document clusters and entity graphs.	34
10 P3, applied to the 2008 US presidential election.	36

Summary

Networks engaged in weapons proliferation, terrorism, cyber attacks, clandestine resale of dual-use imports, arms smuggling, and other illicit activities are major threats to national security. Complexity, dynamism, resilience and adaptability make adversarial networks extremely difficult to identify and characterize. Often the *only* way a threat may be detected is via the networks through which it operates.

To address these national security concerns, in October of 2007 Sandia stood up the “Networks Discovery, Characterization and Prediction” project. The “NGC” was a three-year Sandia Grand Challenge LDRD. Its aim was to research, develop, and evaluate techniques to detect, describe, and then predict the future behavior of networks that pertain to Sandia’s national security responsibilities in cyber security and counter proliferation.

The NGC was aggressively multi-disciplinary. Starting from considerable existing Sandia capabilities in scalable computing and advanced analysis algorithms, it drew on staff from nine of Sandia’s centers, across a wide range of disciplines. It engaged research mathematicians, code developers, computer scientists of both software and hardware, cognitive psychologists, experts in user elicitation, a cultural anthropologist, and many and varied end-user intelligence analysts.

Concrete advances that resulted from the NGC include (from a long list):

- Anomaly detection methods capable of spotting an illicit exfiltration event involving only 0.016% of the collected network traffic.
- The integration of linear algebra, computational linguistics, high performance computing and data warehouse architecture design to build a capability for *translation-free* massive-scale multilingual document clustering on an interactive time scale.
- The ability to predict when fervent on-line conversation will spill over into real-world consequences. One example was the examination of blog traffic in the wake of each publication of the “Danish Cartoons”, with accurate predictions as to whether the on-line discussion of the perceived insult to Islam would result in real-world protest and violence.
- Development of new tool evaluation methods permitting objective measurement of the cognitive load induced by an application’s capabilities, user interface, and visualizations.
- Creation of three substantial “prototypes”, end-user tools. Each addressed a specific set of analyst needs and each incorporated and hardened algorithmic advances to date. They not only achieved their intended goals of testing our algorithmic integration, but also served as vehicles for customer education and program development.
- Improved methods and perspectives on the handling on uncertainty, from technical statistical results assessing the significance and likelihood of newly detected network communities to practical experimentation on how human analysts work with, and around, uncertainty.

- The expansion of the Titan informatics computational framework into an enormously diverse set of network and text analysis capabilities, and its adoption by disparate communities inside and outside of Sandia.

In summary, the NGC staff grew to understand the needs of Sandia's analyst community, did basic research on uncertainty in network analysis, researched and evaluated novel analysis algorithms, and implemented that research to address those needs, creating a flexible, interactive capability for intelligence analysis on large data sets.

In the end, we created at Sandia, in support of the nation, the unique capability to answer previously unanswerable questions.

1 Overview: the Problem, the Goals, and the Approach

1.1 The Problem

Networks engaged in weapons proliferation, terrorism, cyber attacks, clandestine resale of dual-use imports, arms smuggling, and other illicit activities are major threats to national security. These adversarial networks in turn rely on legitimate and illegitimate secondary networks for financial, supply chain, communication, recruiting, and fund-raising activities. Complexity, dynamism, resilience and adaptability make adversarial networks extremely difficult to identify and characterize. Often the *only* way a threat may be detected is via the networks through which it operates.

In short, our real adversaries are networks.

1.2 The Goals

Our goal, then, was to research and develop analysis capabilities that address adversarial networks. The full title of the project, “Network Discovery, Characterization, and Prediction”, conveys the scope and challenges we addressed. The *discovery* of adversarial networks is immensely difficult in its own right. A network may only reveal itself by the union of its parts. Individual relationships and activities may appear completely benign in isolation. Data relevant to network discovery may come from communications, financial transactions, human intelligence reports, shipment records, cyber events or many other sources. It may be geographically or temporally dispersed. Thus, very large and heterogeneous data collections must be analyzed collectively to detect networks. The *characterization* of networks requires methods for identifying likely relationships that are not explicitly captured in the data. The structure of a network conveys information about its purpose and the roles of its component individuals, organizations and activities. It can reveal command and control structure, critical components, and anomalies that are recognizable as such only with reference to the entire network. Structure can also suggest how the network functions, and its likely evolution, allowing *prediction* of the possible future shapes of the network and its real world consequences.

The project team assembled for this Grand Challenge LDRD included research mathematicians, developers, computer scientists of both software and hardware, experts in user elicitation, and end-user intelligence analysts. Building upon considerable existing Sandia capabilities in scalable computing and advanced analysis algorithms, we grew to understand the needs of the intelligence community, did basic research on uncertainty in network analysis, researched and evaluated novel analysis algorithms, and implemented that research to address those needs, creating a flexible, interactive capability for intelligence analysis on large data sets.

In sum, we sought to create at Sandia, in support of the nation, the unique capability to answer previously unanswerable questions.

1.3 The Project Principles

Four persistent themes drove our research and development agenda: analyst needs, semantic graphs, interactivity and scalability, and prototypes.

Analysts Needs Are Central It was our conviction that the analyst is the center of the intelligence universe. Nothing can replace the creativity, judgment and perspective of the human expert. Our goal in this project was not to replace the analyst, but rather to provide unique decision support capability, enabling the analyst to be dramatically more effective. We addressed this not only through innovative research, but through explicitly devoting resources to “human factors” and detailed, involved engagement with our Sandia analyst community, to be sure that we were providing relevant decision support tools in a usable manner.

Semantic Graphs It was necessary that our research directly address missions as diverse as cybersecurity and nonproliferation. Unifying so broad a set of applications and data types was our first challenge, one we resolved by adopting *semantic graphs* as our central underlying data abstraction. In a semantic graph, vertices are nouns (e.g., people, computers, organizations, places) and edges are relationships (e.g., met with, sent packets to, works for). Widely varying types of data can be condensed to sets of entities and relationships, and therefore to the vertices and edges of semantic graphs. Note that different analysis questions will require different data representations; semantic graphs provide that flexibility, in that different graphs providing varying perspectives can be extracted, on demand, from the same underlying relational database.

Further, the focus on semantic graphs allowed the NGC to apply and extend Sandia’s existing expertise in the nuance of multi-way linear algebra. That is, a flat graph with n nodes is mathematically equivalent to an $n \times n$ matrix that indicates whether nodes i and j are connected. A semantic graph is simply a stack of such matrices, one per edge type. That stack is a “tensor”, and so tensor analysis and related linear algebra methods were a constant point of focus, and led in surprising directions: in particular, to advances in text analysis.

Scalability and Interactivity An effective system needs to respond in analyst time. That is, the answer to a question must arrive in time for the analyst to still remember why it was asked. To truly explore data, the response time must be seconds, not minutes, so the analyst can follow hunches and pursue conjectures without losing the overall thread. High performance computing is the only way to provide this degree of interactivity on very large data sets, and this realization was the motivation for the scalability and parallelism activities in the Networks Grand Challenge.

A second aspect of interactive exploration is the need for an assortment of interconnected analysis approaches. A graph algorithm might identify a cluster; an algebraic method could then suggest possible missing relationships within this cluster; and a statistical technique could quantify the likelihood of these putative relationships. This interplay between methods creates opportunities for exploration and discovery that isolated tools can never provide. This is one of the reasons the

NGC focused so heavily on development of the Titan software informatics framework, as a single implementation framework naturally provides the necessary interplay between analysis methods.

Prototypes The fourth of our organizing themes, the use of prototypes, is particularly important in that the role of the prototypes was both so crucial and so easily misconstrued.

The NGC was organized around a series of prototypes with steadily increasing capability. While these prototypes were not themselves the purpose of this project, they did serve to harvest, harden, and instantiate the fruits of the research teams. They were also *targeted* prototypes, developed with the active involvement of a changing roster of analysts, and so designed to address a specific set of use cases pertinent to those analysts.

In the first year we designed, built, and demonstrated the “Thin Line Prototype”, eventually also known as Prototype 1 (P1). P1 addressed cyber forensics problems such as detecting anomalies in port/protocol use or noting whether Internet traffic with worrisome content is crossing interesting political borders. In the second year, we designed and built the imaginatively named “Prototype 2” (P2), which focused on the needs of counter proliferation analysts to master daunting volumes of text. The third year was devoted to a prototype that extended text *and* network analysis capabilities developed in the NGC to provide early warning, from networks of on-line conversation, of topics that will “go viral” and spill over into real world effects, such as violent street protests.

These prototypes served a number of functions. They focused and connected the research threads developed in the NGC in a fashion that let us assess that research for its usefulness; nothing so sharpens criticism as a concrete prototype. They also, and in exactly the same way, allowed us to acquire concrete user feedback on the user interface mechanisms provided by Titan, so that they could may be continuously improved and adapted to analyst mindsets.

The prototypes and Titan-built demos also allowed us to run empirical tests on theoretical claims; around scaling, the accuracy of the predictive topic assessment models, and such not. They built confidence that future, program funded, development focused efforts specifically designed to make use of the NGC’s informatics algorithms to build tools for our analyst community could successfully do so, using the NGC algorithms and their Titan implementation.

The prototypes were very successful in serving the above functions. In fact, they were so successful, and garnered so much attention, that it is probably worth repeating that they were *not* intended to be, in and of themselves, polished end-user applications, an advance in the state of the art, or even necessarily dramatically useful. Further, the prototypes were not our only mechanism for testing our capabilities. Where appropriate, we also engaged in one-off deployments and demonstrations, such as in linking Red Storm to Netezza data appliances for interactive clustering of multi-lingual text data, or the massively parallel topic analysis of all published PubMed papers.

1.4 The Project Structure

Any project this large requires at least some internal structure, as well as some external oversight. The structure was as follows:

- Management Team:
 - Principal Investigator: Bruce Hendrickson, 1410 (FY08), Philip Kegelmeyer, 8900 (FY09, FY10).
 - Project Manager, Bill Cook, 9530.
 - Deputy PM: Suzanne Rountree, 1415 (FY08, FY09), David Rogers, 1424 (FY10)
 - 5900 Liaison: Mark Foehse, 5925 (FY08), Rick Contreras, 5925 (FY09, FY10).
 - 5600 Liaison: Curtis Johnson, 5635 (FY09, FY10)
 - Business intelligence: Mark Huey, Perspectives Inc.
- Our External Advisory Board:
 - Sallie Keller, Chair. William and Stephanie Sick Dean of the George R. Brown School of Engineering and Professor of Statistics, Rice University
 - Matthew Gaston, Senior Research Scientist, Viz, General Dynamics Corporation
 - John R. Gilbert, Professor, University of California at Santa Barbara
 - Karl Kowallis, Network Analyst, Information Operations Center, United States Government
 - Craig Searls, United States Government
 - Peter Weinberger, Senior Software Engineer, Google, Inc.
 - Gerald “Chip” Willard, Technical Director, Office of Analysis, NTOC, United States Government
 - Tom Donahue, United States Government
 - David Jensen, Associate Professor of Computer Science Director of the Knowledge Discovery Laboratory University of Massachusetts Amherst
 - Christopher R. Johnson, Director, Scientific Computing and Imaging Institute Distinguished Professor, School of Computing University of Utah (FY08)

Note that our EAB, by design, was split about evenly between academics in our domain and knowledgeable customers and users. We hosted our EAB for four on-site multiple-day meetings, each meeting culminating in a written report from the EAB laying out their comments, criticisms, and recommendations.

- The Technical Teams and their leads:
 - *Discovery*, Brett Bader, 1415

- *Forecasting*, Scott Mitchell, 1415
 - *Data*, Sue Medeiros, 9512
 - *Integration*, Tim Shead, 1424
 - *Human Factors*, Laura McNamara, 1433
 - *Prototypes*, Laura McNamara, 1433 (FY08, FY09) and Brian Wylie, 1424.
 - *Processor Architectures*, Rich Murphy, 1422
 - *Data Systems Architectures*, Ron Oldfield, 1422
- The **Discovery Team** did fundamental research in graph algorithms and algebraic methods to, loosely, develop and implement scalable methods for the discovery of relevant phenomena in network and text data.
 - Given those discoveries, the **Forecasting Team** addressed “what next, and how do we know?”. That is, it researched methods, necessarily statistical, for the characterization of a network’s properties *and* the uncertainty inherent in that characterization. It also developed algorithms for predicting the future properties of a network as it evolves over time or responds to external events.
 - The **Data Team** wrangled the data necessary for the concrete development and evaluation of the Discovery and Forecasting research. It acquired, ingested, processed, and supplied data sets; some in anticipated support of general NGC research needs, and some on demand, targeted at specific analyst or researcher requirements.
 - The **Integration Team** designed and built the abstractions and mechanisms which implemented the algorithms produced by the Discovery and Forecasting teams, and it built the user interface toolkit used to convert these algorithms into interactive analysis tools. The Integration team, and its consistent focus on implementation within the Titan informatics framework, was the core engine by which the NGC generated a concrete, enduring analytic capability at Sandia.
 - The **Human Factors Team** brought skills in cultural anthropology, cognitive psychology and human/computer interface design to the process of understanding and bridging the actual needs of the analysts to the actual capabilities of the researchers. Further, the Human Factors team evaluated our prototypes and their user interfaces, in partnership with real-world analysts, to return critical feedback to the research and integration teams as to the effectiveness of their work for analyst decision support.
 - The **Prototype Team** was responsible for the design, implementation, and delivery of our software prototypes. It consisted of dual-hatted members of the Integration and Human Factors teams, paired with NGC-funded analysts (different for each prototype) who served as full partners in ensuring that the prototype addressed real analyst needs.
 - In FY09 only, the LDRD office directed additional funding specifically to investigate computer architecture issues pertinent to the NGC. Half of that funding was devoted to the **Data**

Systems Architecture Team, which developed and implemented methods to efficiently integrate graph analysis codes running on an HPC platform with a Data Warehouse Appliance. That is, given existing architectures, it investigated which informatics algorithm should be run where.

- The other half of the funding was invested in the **Processor Architecture Team**, which, conversely, went beyond existing architectures to investigate, simulate, and prototype the optimal architecture for informatics algorithms.

Another way to think about the role of the various NGC sub-teams is to consider how they fit into an overall spiral development model:

1. Conduct fundamental research on analysis methods, algorithms, and architectures. (Discovery, Forecasting, Architectures, Data)
2. Instantiate the fruits of that research into serial and then parallel software tools. (Integration, Data Systems Architectures)
3. Combine capabilities into application-targeted prototypes. (Integration, Human Factors)
4. Evaluate the functionality and utility of these prototypes from the perspective of analysts. (Human Factors, Data)
5. Use the lessons learned to refine the research and development agenda. (All)

There is obviously a great deal of overlap (in staff, objectives, and methods) between the various technical teams, so it should be understood that the technical team structure above was simply a useful mechanism by which to manage and coordinate a project as large and multifaceted as the Networks Grand Challenge.

In fact, to encourage crosstalk and avoid the danger of stovepiping, the research team leads attended each other's technical meetings, all team leads swapped perspectives and activities weekly at a Leadership team meeting, and the PI attempted to attend *all* of the individual team meetings.

2 Activities and Accomplishments

2.1 Advances in Network Analysis

The first of the “project principles” discussed earlier was that “Analyst Needs are Central”. So, though the NGC was primarily a *research* activity, in the set of accomplishment highlights to follow we will be emphasizing the applications as much as the fundamental technical advances that make them possible.

Fundamental Network Properties The NGC developed (and proved) new, fundamental, and practically useful understandings of network properties. These include:

- Methods for hypothesis testing on networks that can help explain to an analyst why the communities detected in a social network are what they are. These derive from fundamental new results in improving the resolution of community detection in networks[BerryHLP] (Figure 1) and analysis of Cohen’s algorithm for triangle finding in networks[BeNo10].

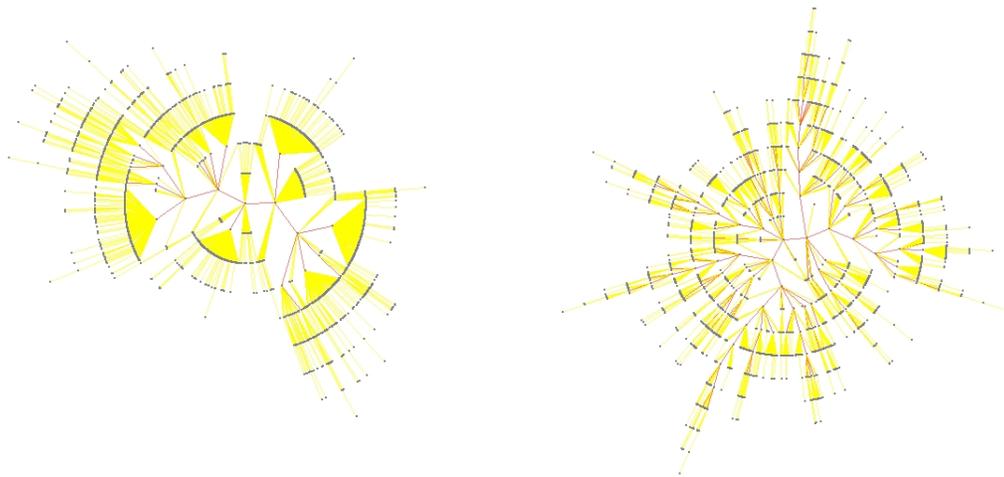


Figure 1. Dendrogram depictions of node partitioning to define communities. The left illustrates the standard CNM partitioning, generating communities which are too few and too big. The right shows the improved wCNM partitioning, which is much deeper, resolving smaller communities.

- Another successful application of hypothesis testing in networks, applied to identification of root causes of diffusely-observed phenomena over time, resulted in finding faulty hardware/software pairs in the Red Sky supercomputer.

- A method for the generation of graphs with specific degree sequences[CaHa09], which is important for testing and validating network analysis algorithms.
- The ability to detect whether a network has a property (known as a “Braess-like paradox”) whereby *adding* connections to improve the topology can actually degrade performance[LeLa09].

In addition to these specific technical accomplishments, our fundamental research efforts resulted in such honors and awards as multiple invitations to give plenary talks (e.g., “Parallel Network Analysis” at SIAM 2009, “Graph Analysis with High Performance Computing” at SIAM 2010) and substantial external funding for spin-off projects initiated by NGC research (“Scalable Methods for Representing, Characterizing, and Generating Large Graphs”, from the DOE ASCR program.)

Novel Tensor/Network Applications We applied tensor and matrix decompositions to cyber and text problems in novel ways, and demonstrated them through, for instance:

- Analysis of computer traffic payload and metadata to find and present related concepts as they develop over time, and analysis of email traffic to discover and understand non-obvious relationships between senders, receivers, time, and topic, such as the effective leadership hierarchy[BaBe09]; see Figure 2.

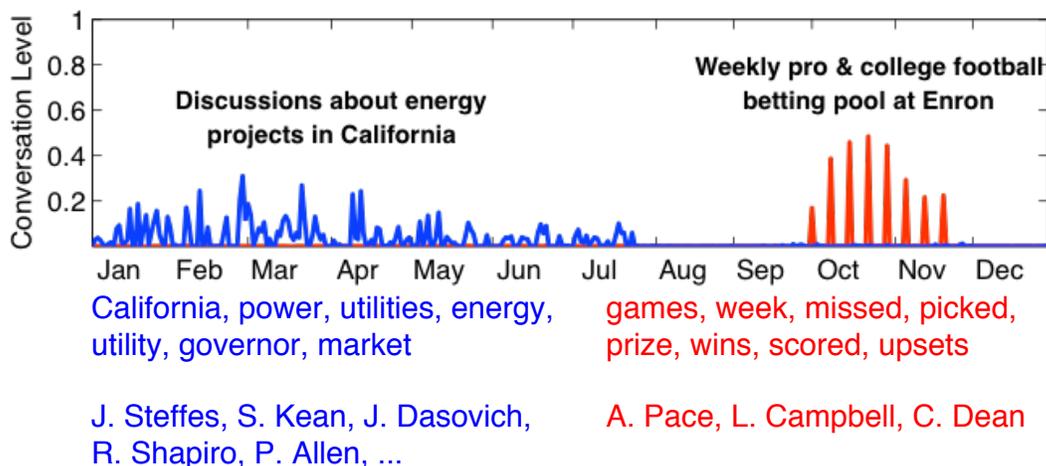


Figure 2. Tensor analysis of the Enron mail corpus finds unusual activity by associating terms with people over time; e.g. Enron’s suspicious energy trading and an unexpected gambling pool.

- Analysis of term adjacency graphs to extract parts of speech from text without the need for prior training on “ground truth”[ChBa09].

- Improved retrieval of pertinent documents from a large corpus through new term weighting schemes for Latent Dirichlet Allocation (LDA)[[WiCh10](#)].
- The integration of linear algebra, computational linguistics, high performance computing and data warehouse architecture design to build a capability for translation-free massive-scale multilingual document clustering in analyst time[[BaCh10](#)], [[OIBaChCCIM10](#)], [[OIBaChSiam10](#)].

Subtle and Scalable Statistical Analysis We developed and demonstrated scalable applications of statistical methods for network and text analysis. These have been applied to:

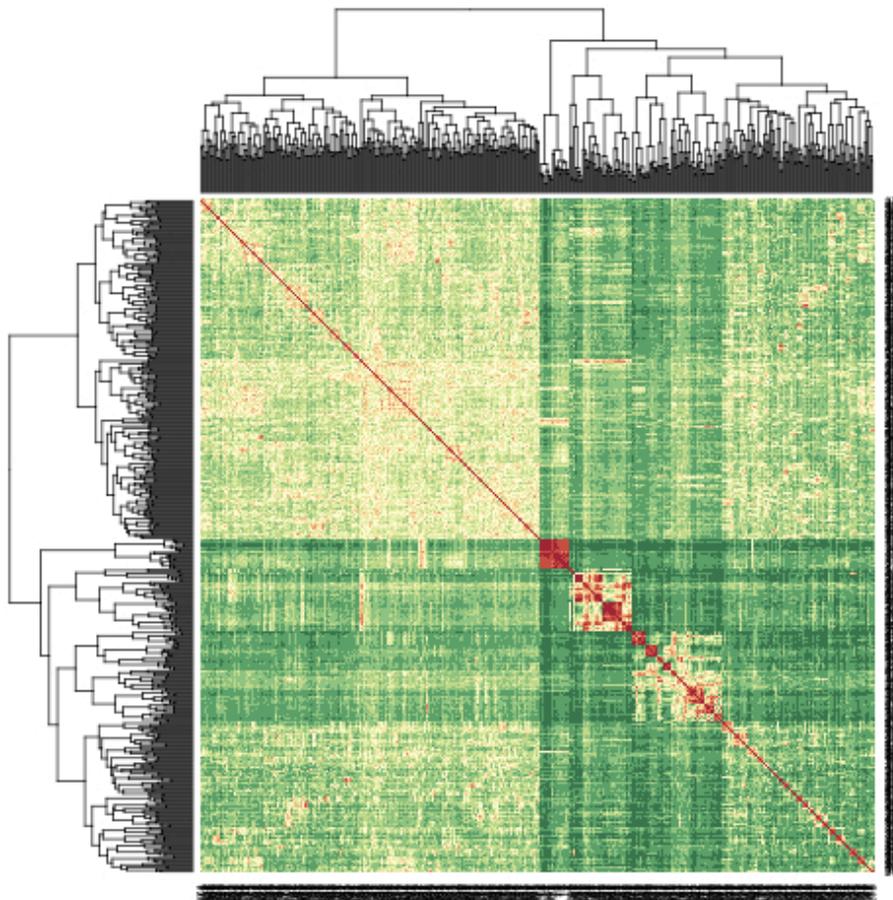


Figure 3. The bright red square indicates the correct detection of 18 anomalous network events, out of a total of 115K events, via LDA statistical analysis.

- The detection of an illicit exfiltration event involving only 0.016% of the collected network traffic[Ro10, Ro10NIPS]; see Figure 3.
- The statistical characterization of communities in email traffic, permitting the detection of *uncharacteristic* mails that may be indications of phishing attacks.
- The simultaneous analysis of both topic and sentiment in text, demonstrated on the temporal development of discussion on political blogs.

Methods for Network Prediction We developed, provided a theoretical basis for, and demonstrated the ability to make *predictive* claims about network behavior:

- The most dramatic example is the ability to predict when fervent on-line conversation will spill over into real-world consequences, as in Figure 4. This work stemmed from fundamental advances in theory[CoGIISI09], [CoGIJMS] and resulted in a number of concrete applications and demonstrations[CoGIISI10], [CoGIASA10], [CoGLOr10]. One such was the ability to examine blog traffic in the wake of each publication of the “Danish Cartoons” and to accurately predict whether the on-line discussion of the perceived insult to Islam would result in real-world protest and violence.
- Other advances were methods for predictive analysis of co-evolving dynamics, for estimating sentiment orientation in social media[CoGlb10], and for detecting emerging topics via meme analysis.
- As a core enabling capability for all of this predictive work, we developed and published some fundamental insights into what makes networks predictable (if they are), how to detect that property, and what the implications might be for intelligence applications[LaGICo09], [CoGoGI10], [CoGla10], [CoGIPC10]. Additionally, we made theoretical and applied advances in simplified but property-preserving analysis of complex networks[CoGIMTNS10], [CoGIISRD].

Scalable Software Infrastructure We developed and implemented libraries for parallel multi-linear algebra for large-scale tensor decompositions, and for parallel statistical characterization, allowing the use of applications at scale[SeBaKo10].

2.2 Engaging with Analysts Via Human Factors

As mentioned in Section 1.3, one of our goals was to to develop prototype information analysis and visualization software tools for intelligence analysts. As these tools were intended to enhance the information exploration and reasoning activities of *real-world* users, the NGC formed a “Human Factors” team charged with supporting the design and evaluation of software from the users’ perspective. However, it quickly became clear that *how* to do such evaluations, in the context of

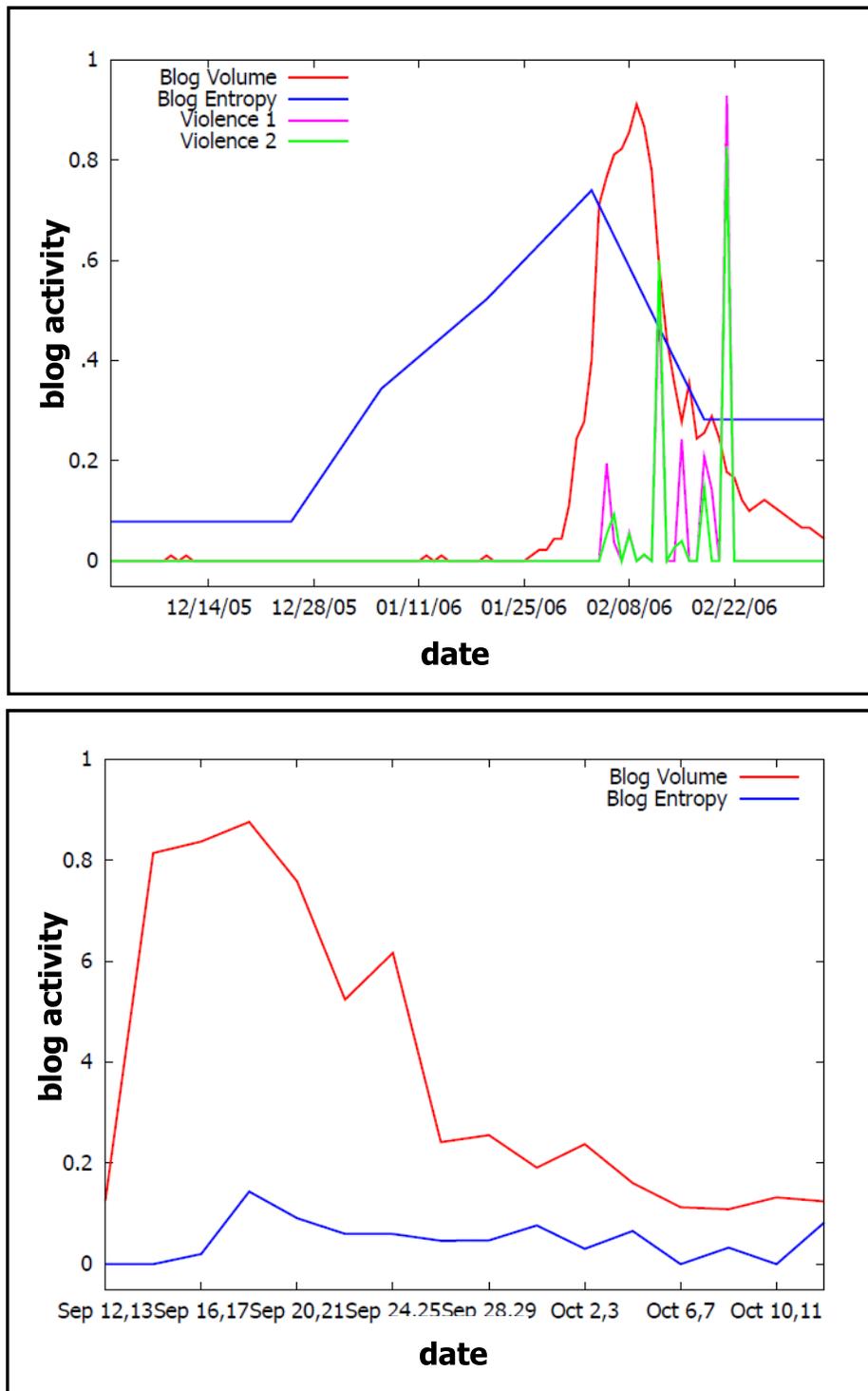


Figure 4. The upper plot, an analysis of the dispersion of blog discussion in the wake of one publication of the Danish cartoons, shows that a peak in entropy *leading* the peak in volume preceded real-world violence. In comparison, the lower plot is a similar analysis of blog discussion in the wake of Pope's speech; the entropy/volume relationship here correctly predicts a lack of real-world violence.

information visualization and visual analytics (Infovis/VA) software, was itself an open question requiring new research.

Evaluation of Visual Analytics Tools

The NGC pursued a number of studies related to the evaluation of visual analytics tools. Most were organized around the NGC's second prototype, P2, which incorporated a number of text analysis and visualization algorithms into the Titan framework to provide capabilities such as document clustering, within-document searching, entity extraction, across-document searches to identify sets of documents that include the same entities, and graphical representations of conceptual relationships across documents; see Section 2.5 for more details. Our studies thus focused on evaluating those functions:

Competitive Testing This study focused on search capabilities. P2 supports a number of search modes, including basic keyword search both within a document and across a corpus and a conceptual search that allows the user to identify a document of interest and search for similar documents using extracted text. We designed a study that compared a commercial desktop search engine, DTSearch, which uses a basic Boolean search strategy to find information in a document corpus.

In this study, we used a between-groups design in which each of the seven participants used one of the two software packages to complete a set of five tasks. All participants received equal training, and were evaluated on task completion, task time, and task load (as computed from the NASA TLX questionnaire).

Due to the necessarily small sample size, we were unable to assert statistically significant differences between P2 and DTSearch. But we did extract results that seemed intuitively satisfying and which were fed back to the NGC Titan and Prototype teams to influence future designs.

Perhaps even more importantly, this study led us to believe that competitive testing is a viable methodology for examining differences in the ways that software supports informatics tasks.

Interaction Design Assessment of P2 . A second study gathered data from an "Interaction Design" perspective. Interaction design focuses on the capabilities the system provides, the interaction space it defines, and users' activities and experiences in using the tool. This study explored these themes in the context of the P2 software. See the separate report [[StHF10](#)] for the study design details and the full set of conclusions, but there were some clear summary observations concerning:

1. The difficulties users experienced in forming mental models of P2 functionality.
2. The utility of giving analysts more interactive control over the construction of clusters and networks.
3. The need to consider the broader context and lifecycle of intelligence analysis in tool design.

The first was more or less expected; though P2's *interface* and affordances had been designed to be very, very simple, nonetheless it was serving up the results of some novel and complicated analysis, and we expected difficulties in conveying that capability. The second point was unanticipated and interesting, and echos the result of the clustering study discussed below: people build mental models by manipulation, and so the more tools you give them for manipulation, the richer the model they can build.

The third point is certainly true, but it was surprising to have it arise here, as the NGC *did* "consider the broader context and lifecycle of intelligence analysis in tool design". Or so we thought. P2's design was a joint effort of not just the Human Factors and Titan development expertise in the NGC, but also that of two end-user analysts. Further, it was informed by consideration of UI and usability lessons gleaned from the first prototype, P1. So it is interesting, and just underscores the difficulty and resource-heavy nature of good interface design, that this third point was still a major source of feedback, despite all the up-front effort directed to UI and workflow,

Investigation of Sternberg Tasks for Informatics Assessment We examined the use of working memory-based techniques, specifically a "Sternberg task", to assess cognitive workload in software user interfaces. The core idea behind a Sternberg task is that, while engaged in some primary activity, participants are *also* presented with three letters that they are asked to remember. As they are performing the primary task, a series of random letters is read to them. When the participants hear one of the letters they been asked to remember, the participant responds by pressing a button. The presumption is that decreased accuracy on then Sternberg task indicates increased cognitive load in the primary task.

Part of this work we were able to do on a separately funded spin-off research project, "Working Memory-Based Metrics for User Interface Evaluation". That investigation used a counterbalanced, within-subjects study, in which participants had to perform a series of simple video annotation tasks using a video viewer tool. One version of the video viewer was designed to be simple and straightforward, while the other version was purposely designed to be confusing and difficult to navigate. Thus we knew, a priori, when cognitive load was increased. And indeed, Sternberg accuracy behaved as predicted, and was even correlated with the subjective NASA TLX survey measure. Thus this study gave us confidence that the Sternberg task can be useful in detecting increased cognitive workload across different interface designs.

Graphs, Cognitive Workload, and Sternberg Tasks The P2 software supports the analysis and clustering of documents in a data set, and displays the output in two types of graph representations: a vertex-edge graph and a ring graph. Examples are in Figures 5 and 6. The vertex-edge graph is the classic "node-and-links" graph. In the vertex-edge graph, individual documents in a data set are represented as labeled and colored nodes. Communities are represented by color at the vertex. The ring graph displays the same information about a data set, but in a set of two nested rings. The exterior ring identifies communities, while interior ring displays all the community members as colored "wedges."

Each of these graph representations is suited to particular tasks, and not well suited to others. Specifically, the vertex-edge graph is best suited to finding relationships between individual

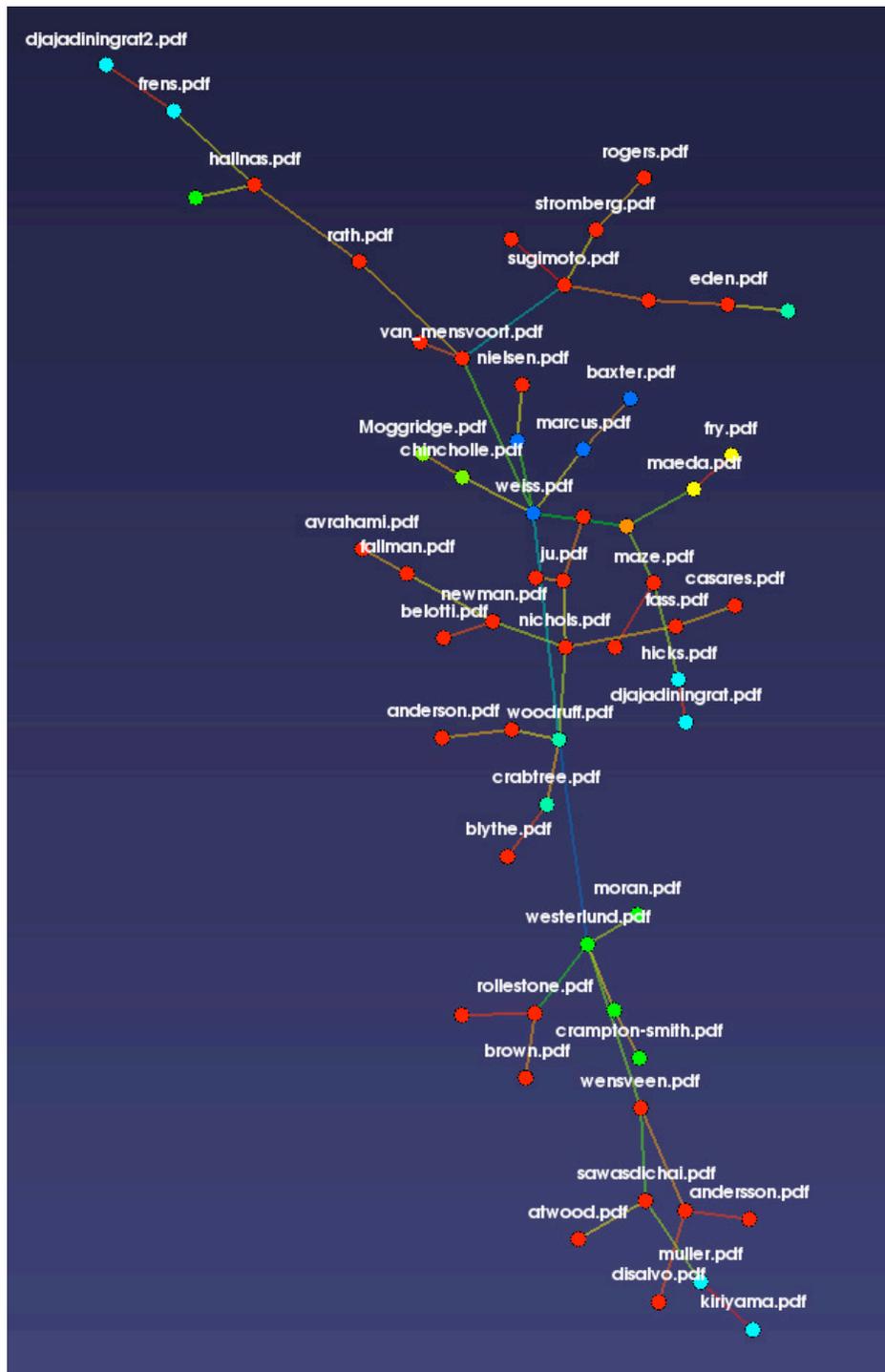


Figure 5. Example of a vertex edge graph

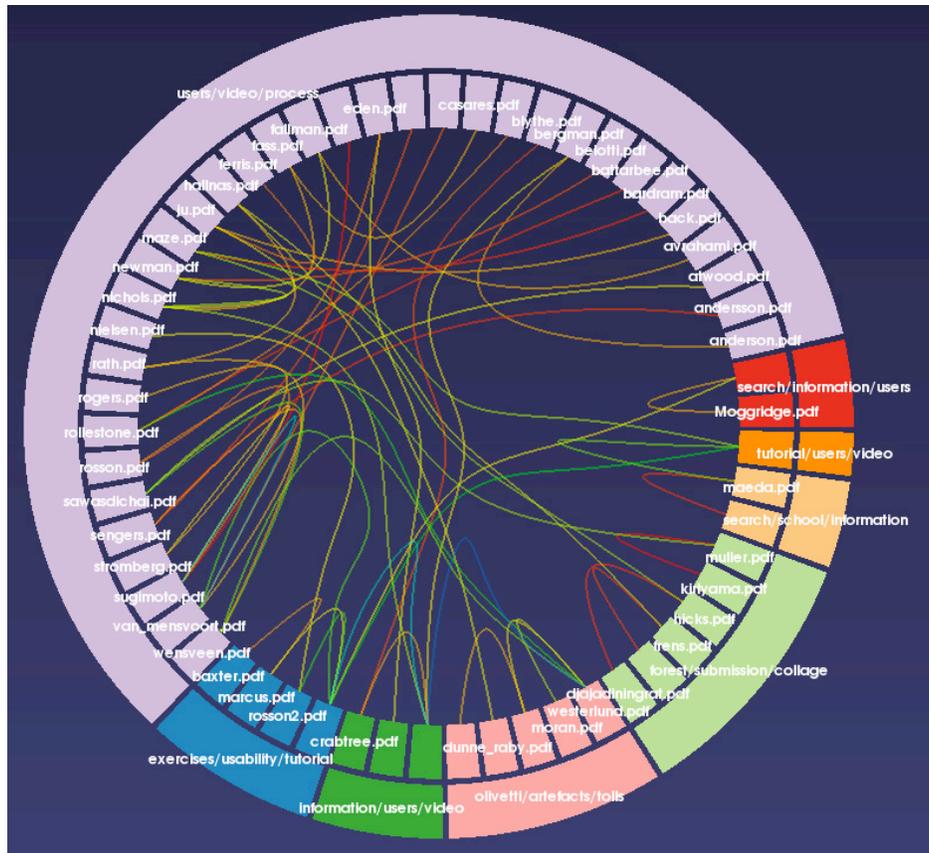


Figure 6. Example of a ring graph

elements (e.g., “What is the shortest path between X and Y?”) while the ring graph is suited to characterizing higher level features of a data set (e.g., “How many thematic communities are in this data set?”). When one of these graphs is used to solve a problem for which it is not well-suited (e.g., using a ring graph to identify a shortest path between individual elements), we hypothesized that the user’s cognitive resources will be stressed, and that their increased cognitive load will be apparent in their performance on a concurrent working memory task.

To test this hypothesis, we designed a 30-participant within-subjects research study to quantitatively compare how people perform on the Sternberg while completing tasks using each of the graph representations. Participants were scored on time on task and error rates for the primary graph tasks, *and* on Sternberg response time and accuracy. Preliminary results suggest that the Sternberg analysis supports the initial hypotheses, that is, that graph task performance would decrease with graph size, and would decrease more when the graph visualization is mismatched to the task question.

Evaluation of Analytic Practice

In addition to our evaluation efforts, we also investigated analytic work practices themselves. In particular, two observations led us to pursue a small document clustering study with analysts.

- When P2 was first deployed in 2009, several of the analysts and the NGC staff members commented that the document clusters were not particularly intuitive.
- Secondly, the NGC researchers were very interested in understanding how analysts assess the quality of information in the source documents they use in their work. In our discussions with analysts, we recognized that the assessment of uncertainty is contextual; as analysts work with collections of documents, they develop a sense for the quality of individual items of information.

These two observations led us to the idea of conducting a clustering study to examine how analysts assess documents and develop topical categories *naturally*, independent of any particular NGC tool. For this study, we worked with a FIE analyst to develop a realistic, hypothetical intelligence question. We then identified 22 intelligence information reports (IIRs) that addressed the question, and that were at no higher than a Secret/National Security Information (SNSI) level of classification, to minimize issues with classification and security. We then recruited 12 intelligence analysts to participate in the study.

The study had four parts: first, a brief interview about their educational and work experience and subject matter expertise, then presentation of the hypothetical intelligence question. The analysts were asked to review the 22 documents, which we shuffled to minimize presentation bias; and asked them to assume that they would be using the documents to develop a report for a customer. We asked them to sort the documents into any categories they felt were appropriate, then debriefed the analysts on their work and categorization strategies. We timed the analysts as they worked with

the documents, noting how many minutes they spent on each document as they were reviewing them. We also took notes on their working strategy and their sorting process as it evolved.

This data will receive continued analysis, but preliminary observations available at the end of the NGC are:

- Firstly, each of the analysts had highly particular categorization strategies. One of the analysts sorted by relevance, another by topic; one ordered the documents by the date-time stamp on the IIR, then re-sorted the documents by topic; another analyst organized the IIRs by the source agency.
- Secondly, the information fields that the analysts used to sort the documents varied as well; some focused on the subject header, while others carefully reviewed the content of the information; still others used header information fields to identify tasking and source information, and used this information to frame the content of the document.
- Thirdly, the analysts seemed to be using the tabletop as a cognitive aid, placing documents into a physical order that reflected their evolving understanding of the documents' content and relationships among the documents.

A tentative conclusion suggests the importance of supporting rich interaction with information: interacting with the IIRs in a physical, flexible way (reading, sorting, annotating) allowed analysts to develop mental models of a dataset's content, and to assess gaps, create new questions, identify search terms, and judge the relevance or quality of individual items in relation to the question at hand.

Other Publications and Accomplishments

To help disseminate our work, NGC staff presented our ideas about working memory for visual analytics evaluation at VisWeek in 2009[DoLA09], at ACM SIGCHI BELIV'10[MaMc10], and at EuroVis 2010[MaMcEuro10]. Partly as a result, we were asked to assemble a working group for VisWeek 2011, focusing on visual analytics software evaluation.

We were also asked to help apply our expertise more directly. That is, we were asked to assemble two review teams for the 2010 VAST contest, teams that included several intelligence analysts, human factors experts, and information visualization computer scientists. We reviewed eight software packages for the 2010 VAST contest[VAST10], and as a result, the National Visual Analytics Consortium at PNNL has asked members of the Sandia Human Factors team to assume a leadership role in organizing panels and presentations on information visualization evaluation for the 2011 Visual Analytics Community conference.

In post-NGC activities, the data we have gathered in the 2010 studies will provide the basis for a series of papers that we will be submitting to BELIV'11, EuroVis, and VisWeek conferences in FY2011 and 2012. Lastly, in FY2011, two of the HF team members will be working on a newly

funded project to study analytic teams and software deployment for an agency in the intelligence community.

2.3 Processor and Data Architecture Insights

In its middle year, the NGC received a substantial one-year funding increase from the LDRD office, specifically to look at the NGC's network issues from a computer systems architecture perspective.

Hardware Custom Architectures

Half of the architecture funds were used to explore potential designs for novel custom processors that would efficiently implement multi-threaded architectures for network problems. The most significant technical outcomes from that work were:

- The creation of a power, energy, and heat model of a 3D stacked implementation of a custom processor for executing NGC-like problems. That model was shared with Micron to aid in understanding the power/performance limits of a 3D stacked approach.
- The creation of a parametrized version of the OpenSPARC processor, including a scalable number of threads, which in turn generated new insights into multi-threaded architectures and potential successors to the Cray XMT platform.
- A summary documentation of this work in the *Network Grand Challenge Custom Architecture Task Final Report* [MuShTe09].

The most significant programmatic outcome is that this NGC-supported architecture effort provided ground work and some early technical results that resulted in a Sandia-led consortium successfully winning Phases 1 and 2 of a DARPA "Ultra High Performance Computing" BAA, resulting in \$8.25M over four years to Sandia specifically, and the possibility of much more architecture research funding if Phases 3 and 4 are bid successfully.

Data Architectures

The other half of the architecture funds were used to investigate *data*, rather than processor, issues. That is, given an informatics problem and its data, and given a heterogeneous assortment of high performance computers and "data warehouse appliances" such as those from Netezza, what is the optimal way to deploy the data and the algorithmic processing across all the architectures?

Investigating that question resulted in three main activities and accomplishments:

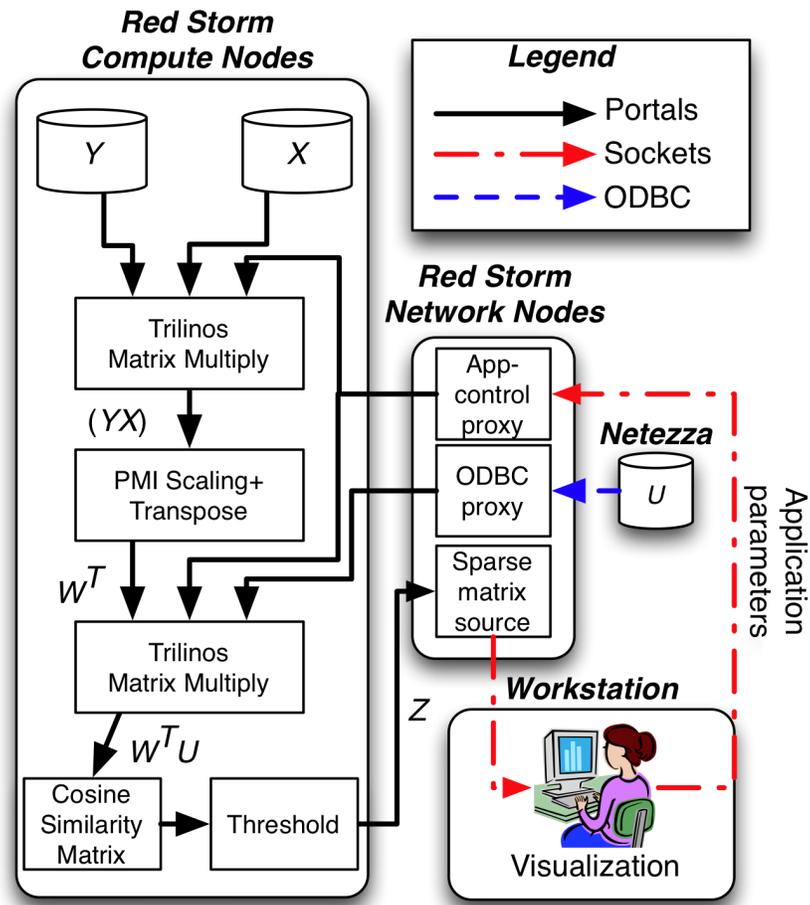


Figure 7. The architectural solution for interactive multi-lingual document clustering connected Red Storm compute and network nodes with an external database appliance machine, all under interactive control of an external workstation.

- *Access to External Resources using Service-Node Proxies.*

One difficulty with integrating HPC and data warehouse appliances is that they didn't, originally, have any way to communicate. So to permit machines such as Sandia's Red Storm to access remote data warehouse appliances such as Netezza, we developed a proxy service that executes on service nodes of the HPC platform. The proxy service communicates with compute nodes using the low-level Portals interface, and it communicates with the database service through the standard ODBC protocol. This approach marked the first time an HPC application could make interactive use of a remote database resource[OIWiDa09]. In addition to database access, this approach has proven useful for interactive visualization and application monitoring, and for real-time streaming of network data for cybersecurity applications.

- *Scaling HPC Resources for Multilingual Document Clustering*

Multilingual document clustering presents a number of interesting challenges that make it particularly well suited for high-performance computing. First, accurate document clustering requires large data sets, both for training and analysis. Second, the computations required to calculate similarities are numerically intensive, requiring large-numbers of matrix multiplies that each take $O(n^2)$ operations, where n is the number of documents in the data set. Finally, time to solution is important in order to support interactive, exploratory investigations. The NGC architecture project developed a large-scale multilingual document clustering application that integrated existing Sandia software libraries including Trilinos, Nessie, and an LDRD-developed code for latent morpho-semantic analysis; see Figure 7. The code demonstrated scalability by clustering the complete European Parliamentary Proceedings (a 1.3 million document data set in eleven languages) in less than 10 minutes on up to 32,768 cores of a Cray XT5[OIBaChSiam10], [OIBaChCCIM10].

- *Integration of HPC and Data Warehouse Appliances*

As an extension to the investigation of the software integration of HPC systems and data warehouse appliances, we explored a more closely coupled hardware integration of the Cray XT3 and XMT with the Netezza data warehouse appliance. This coupling provided a much higher level of performance for applications that have data-intensive characteristics[Old10]. These NGC investigations contributed to initial results that evolved into a follow-on project funded by the ASC/CSRF program.

2.4 Titan as the Enduring Computational Framework

Titan's Role

The Titan Informatics Toolkit [Titan] is a collaborative effort between Sandia National Laboratories and Kitware Inc. It was embraced and thoroughly extended by the NGC, and served as the central computational framework for the NGC's development efforts. NGC scientists were encouraged to do initial brainstorming and algorithm investigation with whatever tools were most familiar

and efficient. But once a given idea proved worthwhile, resources were devoted to implementing the capability in Titan, in order to:

- test and confirm its integration with the rest of the NGC/Titan capabilities,
- to make it available to the rest of the NGC scientists,
- in particular, to make it available for use and demonstration in the NGC prototypes, and
- to ensure that the capability would live on and be usable after the NGC ended.

Overview of Titan

Titan has a number of properties that made it well suited for its role in the NGC. For one, due to its roots in the Visualization Toolkit (VTK) and scalable scientific visualization, Titan provides an excellent framework for doing scalable analysis on distributed memory platforms.

Titan also provides a flexible, component-based pipeline architecture for ingestion, processing, and display of informatics data. Further, it declines to re-invent what it can envelope. That is, it integrates its native capabilities with a series of open-source toolkits for scientific visualization (VTK), graph algorithms (Boost Graph Library and the Multi-Threaded Graph Library), linear algebra (Trilinos), statistics (R), interactive scripting (Python) and more. Titan was one of the first software development efforts to address the merging of scientific visualization and information visualization on a substantive level.

Titan components may be used by application developers using its native C++ API on all popular platforms, or using any of a broad set of language bindings that include Python, Java, TCL, and more. The Python interface has proven particularly attractive to computer scientists who want to make use of Titan's capabilities without having to mount the admittedly substantial C++ API learning curve.

Titan's Development and Accomplishments

Titan has become enormously more capable as an informatics toolkit in the three years of the NGC, both in the depth and breadth of its own technologies, and in the range of informatics application and research it enables. Highlights include:

- The design and implementation of architectures to bridge the gap between large-scale informatics computing and analysts. These included collaborating with the Architecture team on technology to connect HPC resources to external databases (something they aren't designed to do)[[OIWiDa09](#)], the development of a formal understanding of how why informatics applications are different from "normal" scientific simulation and how they can fit within a family of multi-tier client-server designs[[ShTi09-P](#), pages 12–20], and the design and implementation of several variations on multi-tier design[[ShTi10-P](#), pages 14,16,18,22,26].

- Thorough exploration and the development of a mature sense of the design space for creating analyst tools and interfaces. See Section 2.5 for details, but a terse summary is that the prototypes have variously explored the utility, and the consequences, of user interface designs that:
 - expose, but require engaging with, the full expressive power of Titan’s algorithm pipelines (in P1)
 - provide a simple interface targeted at a very focused set of analytic capabilities, at the possible expense of flexibility (in P2)
 - exist as a web interface, rather than a thick client, to make use of browser interface idioms and to separate the backend processing from the user interface (in P3)
- Significant research, implementation, and application around the parallel calculation of statistics in networks[BeGrPe09, BeThPe09, PeTh08, BePeRo09, PeTh09]
- Research, publication, and community outreach on the melding of informatics and scientific visualization[WyBaSh08, CeBaIb08, WyBa09, BaEnOc10, BeAyBa10].
- The development of visualization techniques to address challenges posed by large-scale informatics in both layout[WyBaSh08a] and connected components analysis.
- Making a substantial portion of the NGC technologies available to the broader scientific community as cross-platform, open-source software[ShTi09-P, Slide 5], integrating them into a large set of funded projects at Sandia and elsewhere[ShTi09-P, Slide 7], and in general producing a wide array of useful, tested, and actually used components for serial and parallel informatics ingestion, analysis, and visualization[ShTi09-P, Slide 7].

The NGC’s External Advisory Board has been particularly consistent in its praise for the NGC’s investment in infrastructure in general and Titan in particular. In the third report, two-thirds of the way through the project, the EAB noted:

The NGC has had to invest significant resources in developing “plumbing” to hook everything together, and in the EAB’s view the group has done very well with this work. This infrastructure is something that simply had to be done to enable contributions to the field, and while it may lack the intellectual appeal of the aspect of NGC that has focused on breakthrough research, the Board’s view is that this infrastructure work has been essential for the NGC and will be of high value to Sandia beyond the context of the NGC. The combination of infrastructure and research investments has kept SNL on the cutting edge for the last two years.

And in the final report from the EAB, after noting the “critical advances in Titan”, the EAB go on to say

The contributions to Titan should not be undervalued by Sandia. Sandia has made a large investment in the Titan platform. The NGC impetus has really made Titan come to life and will help Sandia realize a return on their investments.

2.5 The Three Prototypes

As mentioned in Section 1.3, the NGC was organized around a series of prototypes with steadily increasing capability. Each prototype was intended to be a functional, if rough, interactive tool deploying NGC-developed and newly integrated capabilities against specific application use cases.

We will below briefly describe those use cases and the core functionality of each prototype. Still, as interactive tools, seeing the prototypes in action is the best way to appreciate their nature. To that end, demonstration movies were created and are available on line for all three prototypes[NGC].

Prototype 1: Network Communication Analysis

The first of our prototypes, P1 (Figure 8), was also known as the “Thin Line” prototype, as its integration goal was to demonstrate that the NGC/Titan tools could take at least one continuous path, from start to end, through the welter of analysis options being generated by the NGC. P1 operated on network communication data and addressed four cyber forensics use cases elicited from our analyst partners:

- Which machines are talking to each other?
- Which network transfers crossed political boundaries?
- What payloads were contained in the network transfers?
- What out of the norm behaviors should be flagged for further exploration?

Products from the Discovery, Forecasting, and Titan teams incorporated in this prototype included:

- A tensor approach for analyzing the content of any payload text, via the Titan integration of the PARAFAC analytic tool. This analysis takes in source Internet protocol (IP) addresses, textual content, and time stamp data, to reveal related concepts over time in the outgoing IP traffic.
- Graph analysis algorithms including short path discovery, connection subgraphs, and community finding.
- Visualizations specifically designed for algebraic methods.
- The earliest version of a suite of attribute-based statistical metrics.
- Ring-view and 2D geovisualization views, developed in direct response to Sandia analyst feedback.

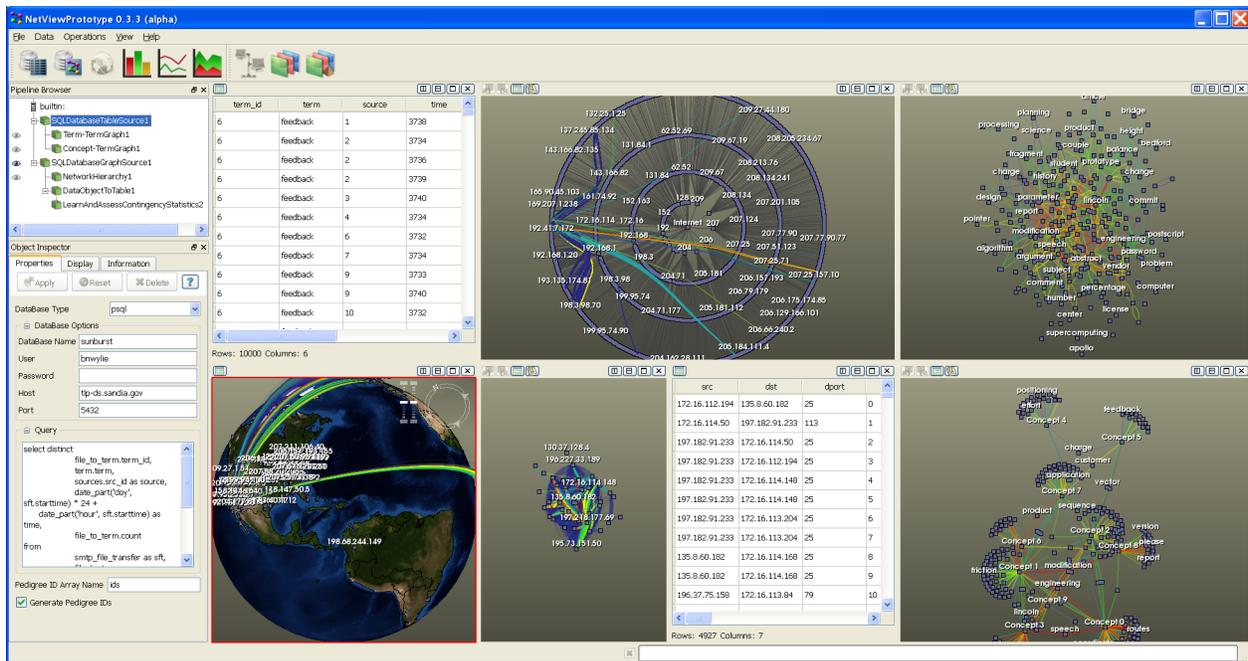


Figure 8. P1, illustrating almost all of its visualizations.

This prototype admirably served the original goals of the prototypes discussed earlier. That is, it proved that Titan’s data flow and filter architecture was indeed suitable for informatics problems, and demonstrated (not just to the NGC staff or to Sandia, but also to the broader intelligence community and potential commercial partners, via repeated live demos) that we can combine the fruits of our research and visualization methods to address practical problems. It also served as a concrete mechanism by which we elicited analyst feedback, on user interface as much as capability, and thereby improved subsequent prototypes and the Titan user interface tools in general.

Even given all that payoff from this prototype effort, it was indeed designed to be a *prototype*, unpolished, intended for use only within the NGC. We were therefore surprised, but pleased, that our analyst partners on P1 found P1 to be useful enough that they, on their own funding, took it back to their systems and used it to explore real cyber forensics data:

“The work developed in this LDRD has a lot of potential. P1 (an application built on the Titan toolkit) has already been valuable in addressing several cyber analysis tasks that were presented to Sandia. I have much confidence that through continued collaboration this work will benefit the Labs and the nation in the area of Threat Analysis and Awareness.”

– Mark Hollingsworth 5631

Prototype 2: Text Analysis for Counter Proliferation

The second prototype addressed a very different problem. Here the goal was to do analysis of unstructured text in support of the information ingestion needs of counter-proliferation analysts.

That is, the usual workflow is that an analysts gets a question from a customer about a topic that's new to her. She conducts iterative searches through multiple databases, reading as she goes along. She pulls documents that seem relevant to the question and stores them on her desktop computer. When she feels she's gotten everything she's going to get out of the databases, she reads and annotates what she's collected and writes the report.

Since no analyst could read and absorb *everything* that might be pertinent, especially in the scant time usually available, the quality of the eventual reports often depends greatly on the analyst's information triage skills. That is, out of a thousand documents turned up in a search, the analyst may read the titles of all, the abstracts or the skimmed bodies of two hundred, and read only fifty in depth; a good report depends on having selected the right subset to focus on.

The nominal goal of P2 was to aid in document organization and absorption so as to enable the analyst, with no additional time required, to write the report she would have written had she actually been given time to read all one thousand of the original documents.

To that end, P2 provided, in one interface (Figure 9), the ability to:

- Ingest documents in a variety of formats, including internal intelligence reports (IIRs).
- Automatically organize those documents topically.
- Extract entities such as person names, places, and organizations, and to present the original documents with those entities color coded and linked to further on-line reference material such as provided by Wikipedia.
- Permit the analyst to interactively build and adjust a "hot list" of entities of interest. In response, P2 would construct, in real time, the smallest network of plausible connections between those entities. One novel analytic capability highlighted here, one not available commercially, is the ability to find subtle connections between entities via document similarity, even when there is no one document that explicitly referenced them both.

P2 was also interesting in that it had, in contrast to P1, a drastically simplified interface. This made it simple enough that users could, and did, download an installer and work with P2 on their own documents, without aid from NGC staff. On the other hand, it turned out that it also constrained the ways a user could operate with, and build mental models of, their data, a point that received considerable human factors attention in the third year of the NGC.

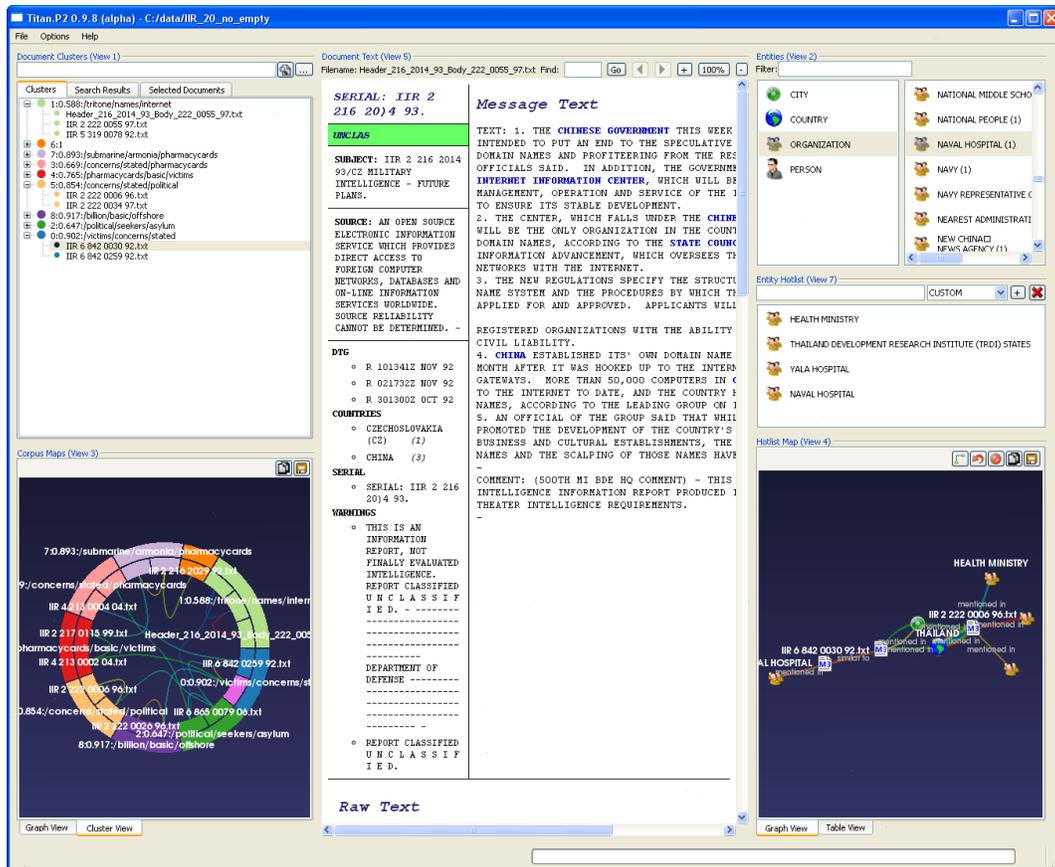


Figure 9. P2, illustrating document clusters and entity graphs.

Prototype 3: Early Detection of Internet Memes

The first prototype explicitly focused on cybersecurity as an application, and on communication networks as a data type. The second prototype focused on counter proliferation as the application domain, and on text, and the entity relationships deducible from text, as the data type.

The ambition in the third prototype was to address, in a single application, a superset of all that had come before, and to add an additional dimension: prediction. The goal was to be able to watch evolving on-line conversation, *multi-lingually* and without the need of translation, to order to do automatic and early detection of core topics, “memes”. Further, P3 was to predict which of the detected memes (and which of any memes suggested by human analysts) will “go viral”. In other words, the idea was to build a multi-lingual early warning system for issues that will spill over from on-line discussion into real-world consequences.

Algorithmically, this was a very nice mix of NGC capabilities. That is, given a set of date stamped blog posts in multiple languages, P3 applied a slew of methods already developed in the NGC:

- The weighted CNM algorithm for finding communities in those blog posts.
- Multi-lingual methods for clustering documents by topic.
- Communication entropy measures derived from stochastic/hybrid networks models that could be used to note, and contrast, the volume and spread of discussion on a given topic.

The NGC had already demonstrated a core part of this capability, predicting the likely impact of analyst-suggested memes, in its second year. That capability existed only as a set of ad hoc desktop researcher tools; so part of the point of P3 was to harden and productize those capabilities, bringing them fully into the Titan fold. As a result, not only was this predictive capacity now in Titan, but the full set of Titan mechanisms for data visualization and analysis could now be deployed against this on-line text data, creating rich interactive tools for investigating the temporal and clustering properties of the memes.

Another point to P3 was to explore a new implementation and delivery mechanism for NGC prototypes. P1 and P2 were thick clients that had to be installed on the user’s desktop. P2 in particular was stripped down and packaged with an installer that made the installation straightforward, but still, making use of every update and improvement to P2 required re-installation. P3 (Figure 10) was developed as a web application, one activated simply by visiting a web page with a browser. This meant that:

- “Upgrades” were automatic, just a matter of revisiting the page,
- Data processing could be made asynchronous, so that processing whose requirements are hard to predict (such as cleaning the nightly web crawls) could essentially be run off-line without interfering with the analysis of the existing data,

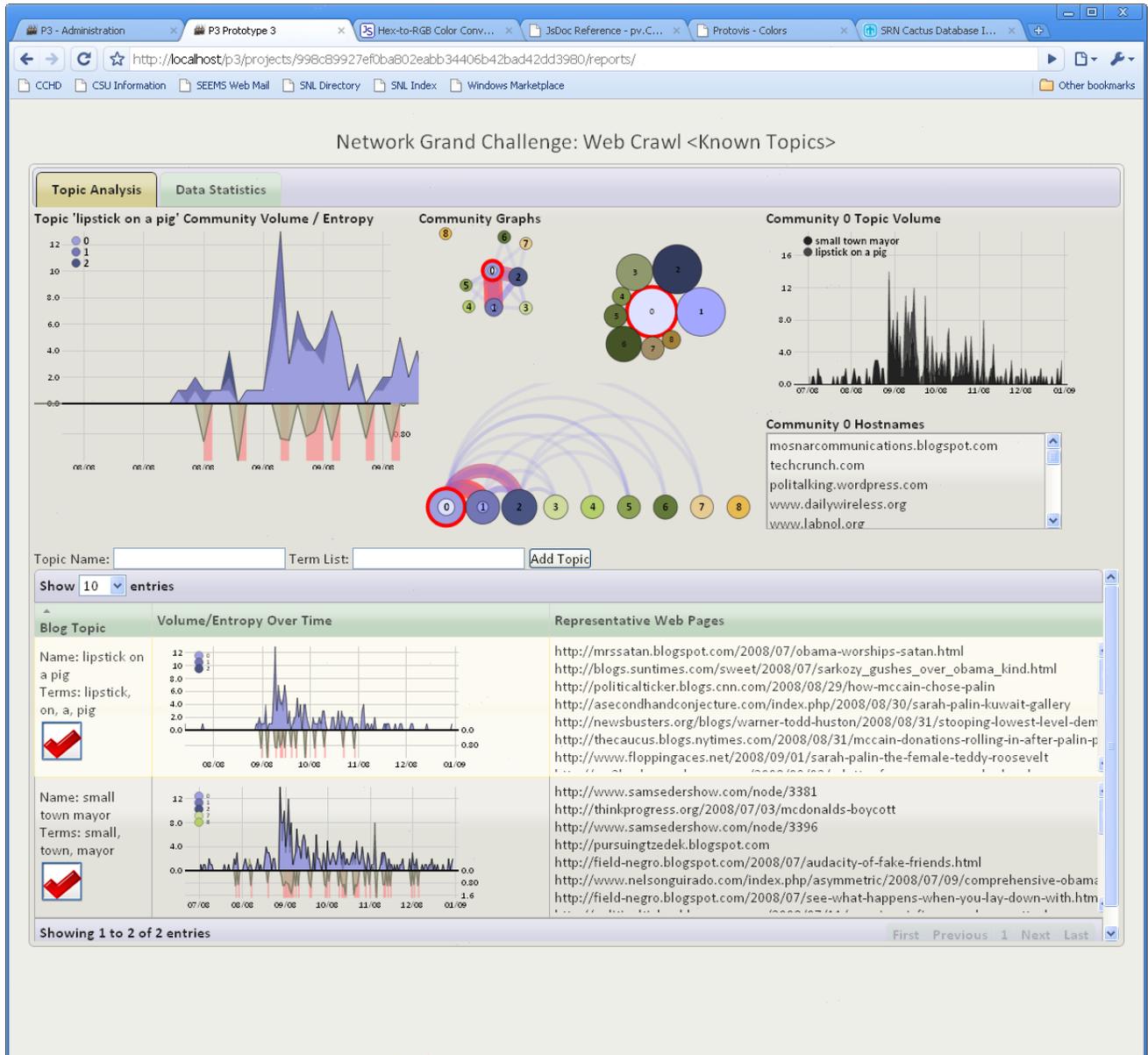


Figure 10. P3, applied to the 2008 US presidential election.

- New UI tools and methods, appropriate for web interfaces, could be investigated and evaluated,
- All without reducing personalization; multiple users could use the same server to explore the same (or different) data sets, and maintain separate saved states and memes of interest, all without interfering with each other.

2.6 Acquiring and Wrestling with Data

It is a truism, too often neglected, that a substantial amount of the work in any data analysis project, even research-oriented projects, is necessarily devoted to data acquisition and preparation. It was expected that the Networks Grand Challenge would be no exception, and so a fair amount of the NGC's resources were devoted to such issues.

Policy and Tools

One of the earliest concerns around data acquisition was to understand the proper practice around privacy issues. To that end the NGC formed a Security and Privacy Oversight (SPOT) Team, consisting of NGC staff, a representative from Sandia legal, and senior managers, to develop a set of guidelines for determining the suitability of data sets for use by the NGC, with proper consideration given to Executive Order 12333, the Privacy Act, and the Data Mining and Reporting Act. These guidelines were featured on the wiki, publicized at All Hands meetings, and consulted by the SPOT team each time they reviewed a new data set proposed by some NGC staff member.

The NGC also made software and hardware investments to support general data processing and serving. These included:

- Two Inxight licenses, for text tokenization, entity extraction, and general pre-processing.
- Multi-terabyte disk storage, purchased or rented from Sandia corporate resources, for data storage and service.
- Acquisition and installation of the "Cactus" web crawling software, for acquisition of blog data.
- A multi-core, multi-terabyte Linux server for third year real-time processing in support of the third prototype.
- Use of Sandia's corporate high performance computing resources, in particular Red Storm, Red Squall, and Red Sky.

In addition, a fair amount of home grown tool development and configuration was required, including:

- Considerable customization and configuration of the Cactus tool and its output, resulting in multiple coding efforts to isolate and extract blog posts from web crawl data, and to determine date stamps reliable enough to fuel the desired temporal analysis.
- Custom high performance loaders of Inxight data for Netezza ingest.
- Custom data base support; benchmarking, tuning, schema design and maintenance.
- Custom data manipulation, such as resolving UTF country codes when preparing multi-lingual data for clustering analysis.

These may seem prosaic activities, but they are critical to high quality data, and that is precisely why they are highlighted here. It is all too easy in the design of informatics projects to underestimate the time and resources required by effective data preparation. The PI, despite having learned this lesson before, was still caught short by these issues a time or two; so it worth dwelling on this point, to serve as a caution for future projects.

A Sampling of Data Sets

Described below are some of the data sets acquired by the Networks Grand Challenge and analyzed by its researchers:

IIRs

468 *unclassified* “internal intelligence reports” (IIRs), acquired through search for such, followed by the derivative classification of the reports, individually and jointly, all before release for use by the NGC staff.

Though this was easily our smallest data set, it was also one of our most unique and most useful ones. The second NGC prototype, in particular, was focused on analysis and organization of the sources a counter-proliferation analyst would assemble for an investigation. IIRs are a primary source, and further, they differ dramatically in format, appearance, metadata, and expressive idiom from most other text data. So the IIR database allowed us to wrestle with those issues up front, and address both the practical and technical issues in applying NGC analytic capabilities to that data.

Meme Tracking: Cactus Crawl

4 Gigabytes of political blog postings, primarily in English, on the theme of the 2008 US presidential elections, with full text and site link information. Generated via Cactus.

This crawl was inspired by the Leskovec and Kleinberg paper “Meme-tracking and the Dynamics of the News Cycle”, which identified fifty memes. The idea is that each of these fifty was a meme that is known to have persisted, rather than dying off, and could thereby be used as “ground truth” to measure our own predictive accuracy in P3.

Danish Cartoons: Cactus Crawl

Two sets of crawls over multi-lingual blog postings, on the theme of the Danish cartoons, with full text and site link information.

PubMed

15 million article citations from PubMed with complete title, author, keyword and publication data. 8.8M have attached abstracts, totaling 15GB of text. With thanks to Juliana Freire, University of Utah, for the data.

The Bible

In several languages, pre-tokenized by white space, 282 MB, 2 million documents.

The Qur'an

In several languages, pre-tokenized by white space, 7 MB, 37K documents.

Europarl

European parliamentary proceedings, translated into all languages of the European Union. 2.2 GB, 4 million documents.

NY Times Annotated Corpus

100 GB XML-formatted NY Times annotated articles, consisting of over 1.5 million articles manually tagged by The New York Times Index Department, with a normalized indexing vocabulary of people, organizations, locations and topic descriptors.

Wikipedia

A snapshot of all English Wikipedia text.

SAND Reports

All unclassified Unlimited Release Sandia technical reports from 1970 until September 2009. In PDF format, 9,335 files, 67 GB.

2.7 Programmatic and Cultural Outcomes

So far this report has focused mainly on the technical activities and accomplishments of the Networks Grand Challenge. Also very important, especially in terms of future and enduring impact, are the programmatic advances and cultural changes enabled by the NGC.

Technical Communication

Part of the research responsibility in any LDRD is engagement with the broader technical and academic community. This was complicated for the NGC by the nature of our application focus,

but nonetheless we were generally able to find or create surrogate data sets or applications and engage externally.

In particular, we have published (to date) fifteen journal publications, eight invited papers and plenary addresses, and thirty-eight conference papers and presentations. These publications, as well as other project summary information, are available at ngc.sandia.gov, a public web site providing convenient access to the openly available fruits of the NGC's labor.

We have also been active in engaging academia through support of summer students and university research. Particularly significant has been collaboration on social network dynamics with Kristin Glass of New Mexico Technical Institute, and on statistical analysis as applied to network uncertainty, with Alyson Wilson and Dan Nordman of Iowa State.

These publishing and communication activities have led to higher level advisory engagements as well. Members of the NGC Human Factors team were asked to join the review staff for future "Visual Analytics Science and Technology" contests, the NGC PI was asked to participate on the External Advisory Board of network analysis efforts at Notre Dame and Lawrence Livermore National Labs, and the PI was also invited to address the JASONS as part of a study investigating the path "From Data to Decisions".

Business Development

One programmatic goal of the NGC was to establish an enduring area of expertise in network analysis, one that would result in continued opportunities for application of existing capabilities to national security problems, and continuing research funding for developing new capabilities.

To that end, the NGC Management and Leadership teams engaged in exhaustive (and sometimes exhausting) communication and outreach throughout the project. We kept track of these contacts in an "Interactions Log"; the record indicates that we conducted a business briefing roughly once a week throughout the life of the project. The targets of these briefings came from partner agencies, commercial firms, academia, the DoD, DHS, DoE and DoS, the national labs, and NIST. Some of the briefings were to internal Sandia management, at all levels, to help craft proposals and inform program development.

As one result, 5600, 5900, 1400, and 8900 have developed the cross-organizational infrastructure and awareness necessary to most effectively mutually support each other in business development. For instance, over the course of the NGC, two additional managers and four additional staff in 1400 received the higher clearances needed to best engage with 5000 problems and data. It has now become the expected norm, as evidenced by the FY11 EPS and Late Start Cyber processes, that research programs purporting to tackle 5000 class problems discuss the specifics of their 5000 engagement as part of the proposal. Similarly, informatics projects in 5000 now routinely reach out to staff in 1400 and 8900 for advice or to fund staff.

It is difficult to claim success in our business ambitions, as that is the sort of judgement best made in retrospect, but early indications are encouraging. Already, at least twenty-five new projects

have been started based on the foundations laid by the NGC, projects that address scalable methods, custom HPC architectures, anomaly detection, social dynamics, and cyber analysis.

Further, the funding brought in by these projects more than matches that invested in the NGC, and, even more encouragingly, half of that funding comes from external sponsors, including partner agencies, defense contractors, the DOE, DoD, DHS, DARPA, the Army Research Lab, and IARPA.

Cultural Integration

Finally, the goal of the NGC was not solely to spend three years assembling and applying technical muscle to one-off solutions of some of Sandia's national security problems. Rather, we started from the understanding that the research community and analyst community we were attempting to connect did indeed need connecting. They knew distantly about each other, but the research staff did not well appreciate the analysts' problems, and the analysts did not well understand what technical capabilities were applicable to their problems.

The NGC's long-term impact, therefore, depends also on how well that gulf has been bridged. We've cited some metrics and figures above, but two anecdotes stand out:

- Mark Foehse is a counter proliferation analyst in 5925. In FY07, he knew none of the staff, and none of the technical capability, in 1400 and 8900. Midway through FY10, asked to lead a short term, high visibility, late start LDRD, he not only understood enough of Sandia's network analysis capability to sensibly scope the project and its deliverables, but he also knew which specific researchers to approach to staff the project.
- In FY07, the Titan framework was a tool known about only within 1420, and used as an application development tool essentially only by its developers. In FY10, there are several 5000 staff dedicated to using Titan to build and modify applications. With the result that, for instance, during a discussion at the final EAB meeting, Michael Stickland (staff in 5635) was making insightful comments about the minutia of the Titan build process.

The EAB summarized the cultural impact in their final report:

The GC effected cultural change at Sandia across various research and analysis sub-cultures ... At the initial EAB meeting, it was apparent that many contributors from the various organizations barely knew each other's names; whereas now contributors are routinely presenting each other's work.

3 Conclusion

3.1 Continuing Challenges and Opportunities

It is inevitable that any limited-duration project with the scale and the ambition of the Networks Grand Challenge will raise as many questions as it addressed. The issues underscored and only partly addressed by the NGC are both technical and cultural:

- *Technical:*

- Uncertainty analysis in the network context remains a challenge. The NGC made point-wise advances, and certainly deepened the understanding of the problem domain, but a general methodology is still elusive.
- One goal of the NGC was to make algorithms effective on simple graphs equally effective on full n-way attributed relational graphs. Much progress was made here, e.g. with tensor analysis and in community detection, but other fronts still require attention, connection subgraphs being an excellent example.
- It remains difficult to deal with data that is missing, noisy, uncertain, or corrupted. Various specific instances of challenging data were tackled and handled by the NGC, but the development of general principles for such problems is likely its own Grand Challenge.

- *Cultural:*

- An ongoing challenge is understanding an analyst’s work processes at the right level to usefully guide algorithm development and software integration. We have made, and published, both practical and theoretical progress against this aim. Still, the experience of intimately involving analysts in P2’s use case and UI, yet later learning that we thereby abstracted away some of the manipulative tools other analysts need to engage their data, clearly indicates the value of further human factors work.
- Similarly, insight and experimentation is needed to understand how explain or deploy sophisticated algorithms in a fashion that allows analysts to trust and use them without having to master the underlying mathematics.
- Finally, even for useful and usable tools, assessing the specific benefit of those tools on analyst effectiveness remains a research question. We’ve suggested a fertile new dimension of working memory tests as a means of measuring cognitive load, but there is still much to do done. Even making the concept of measuring “rate of inspiration” well-posed is a challenge in itself.

3.2 An Outside Perspective

Given the above list of incompletely met challenges, and the difficulty of accurate self-appraisal on such a complex and recent project, perhaps the best final comments might come from our External Advisory Board, in the report stemming from the on-site meeting held a month before the project ended.

- On the NGC's technical accomplishments:

There are numerous areas of technical accomplishments, including algorithmic advances, novel multi-language document clustering, critical advances with Sandia's Titan framework, infrastructure to attack Big Data in varying domains, and groundbreaking human factors work.

...

The contributions to Titan should not be undervalued by Sandia. ... The NGC impetus has really made Titan come to life and will help Sandia realize a return on their investments.

...

The basic and foundational R&D of the NGC has put Sandia in an extremely strong position. Sandia — and even the NGC — may perhaps not yet be aware of what a valuable vein they are mining.

- On our cultural and programmatic contributions:

The GC effected cultural change at Sandia across various research and analysis subcultures ... At the initial EAB meeting, it was apparent that many contributors from the various organizations barely knew each other's names; whereas now contributors are routinely presenting each other's work.

...

There are also notable programmatic successes: traction for various prototypes within the intended user community, significant follow-on research and application funding, and strong potential for additional application work.

- And, finally, on our initial ambitions and overall achievement:

The NGC has gone further than anyone should have expected ... Three years ago the EAB's major concern was that the project was taking on too much ... Now, the EAB finds that the NGC has not just met but exceeded some of the stated goals which the EAB worried were not attainable.

...

There is a real chance for SNL to make a name for itself based on this work, to become the go-to place to provide the infrastructure for some of the biggest trends in informatics and security, including cyber security. The EAB can glimpse a future where Sandia's signature capabilities are traceable to the NGC.

Annotated Bibliography

- [BaBe09] Brett Bader, Michael Berry, and Amy Langville. Nonnegative matrix and tensor factorization for discussion tracking. In Ashok Narain Srivastava, Ashok Srivastava, and Mehran Sahami, editors, *Text mining: classification, clustering, and applications*. Chapman & Hall/CRC, 2009.
- [BaCh10] Brett Bader and Peter Chew. Algebraic techniques for multilingual document clustering. In Michael Berry and Jacob Kogan, editors, *Text Mining: Applications and Theory*. Wiley, 2010. DOI: 10.1002/9780470689646.ch2.
- [BaEnOc10] J. Baumes, A. Enquobahrie, T. O’Connell, T. Otahal, P. Pébay, W. Turner, and M. Williams. Integration of R to VTK, adding statistical computation to a visualization toolkit. Conference Presentation at useR!, July 2010.

Abstract: Conveying the sense of complex data to the human mind requires sophisticated visualization methods. The Titan informatics toolkit, a Sandia funded collaboration between Sandia National Laboratory and Kitware, represents an effort to add graphical, tabular, and geospatial visualization algorithms to the Visualization Toolkit (VTK). VTK is an open-source, freely available software system for 3D computer graphics, image processing and visualization. The Infovis additions to VTK expand the the toolkit to include visualization of spatially ambiguous entities. However, simply displaying relationships among entities is not sufficient. Statistical analysis such as that provided by R is a powerful tool for suppressing noise in the data and enhancing real relationships. This abstract describes the addition of an R interface to the VTK toolkit and introduces the use of the R engine in several VTK Infovis application areas.

- [BeAyBa10] Berk Geveci, Utkarsh Ayachit, Jeffrey Baumes, Michael Bostock, Vadim Ogievetsky, Brian Wylie, Timothy M. Shead, Emanuele Santos, Timo Ropinski, and Jorg-Stefan Prini. DIY Vis applications. Tutorial at VisWeek 2010, October 2010.

Abstract: Every year, researchers present many new wonderful visualization and analysis algorithms. However, many of these algorithms are not transitioned to receptive researchers in a timely manner. Algorithm developers typically build lightweight prototypes to demonstrate their ideas and research to the community, and building full-featured visualization applications is hard work. This tutorial covers some of the most popular open-source frameworks whose aim is to simplify the development and deployment of visualization algorithms to high quality software applications.

- [BeGrPe09] J. Bennett, R. W. Grout, P. Pebay, D. Roe, and D. Thompson. Numerically stable, single-pass, parallel statistics algorithms. In *IEEE Conference on Cluster Computing*, pages 1–8, 2009.

Abstract: Statistical analysis is widely used for countless scientific applications in order to analyze and infer meaning from data. A key challenge of any statistical analysis package aimed at large-scale, distributed data is to address the orthogonal issues of parallel scalability and numerical stability. In this paper we derive a series of formulas that allow for single-pass, yet numerically robust, pairwise parallel and incremental updates of both arbitrary-order centered statistical moments and co-moments. Using these formulas, we have built an open source parallel statistics framework that performs principal component analysis (PCA) in addition to computing descriptive, correlative, and multi-correlative statistics. The results of a scalability study demonstrate numerically stable, near-optimal scalability on up to 128 processes and results are presented in which the statistical framework is used to process large-scale turbulent combustion simulation data with 1500 processes.

- [BeNo10] Jonathan W. Berry, Daniel J. Nordman, Cynthia A. Phillips, and Alyson G. Wilson. Listing triangles in expected linear time on a class of power law graphs. SAND Report SAND2010-4474C, Sandia National Laboratories, 2010.

Abstract: Enumerating triangles (3-cycles) in graphs is a kernel operation for social network analysis. For example, many community detection methods depend upon finding common neighbors of two related entities. We consider Cohen’s simple and elegant solution for listing triangles: give each node a “bucket.” Place each edge into the bucket of its endpoint of lowest degree, breaking ties consistently. Each node then checks each pair of edges in its bucket, testing for the adjacency that would complete that triangle. Cohen presents an informal argument that his algorithm should run well on real graphs. We formalize this argument by providing an analysis for the expected running time on a class of random graphs, including power law graphs. We consider a rigorously defined method for generating a random simple graph, the erased configuration model (ECM). In the ECM each node draws a degree independently from a marginal degree distribution, endpoints pair randomly, and we erase self loops and multi-edges. If the marginal degree distribution has a finite second moment, it follows immediately that Cohen’s algorithm runs in expected linear time. Furthermore, it can still run in expected linear time even when the degree distribution has such a heavy tail that the second moment is not finite. We prove that Cohen’s algorithm runs in expected linear time when the marginal degree distribution has finite $4/3$ moment and no vertex has degree larger than \sqrt{n} . In fact we give the precise asymptotic value of the expected number of edge pairs per bucket. A finite $4/3$ moment is required; if it is unbounded, then so is the number of pairs. The marginal degree distribution of a power law graph has bounded $4/3$ moment when its exponent α is more than $7/3$. Thus for this class of power law graphs, with degree \sqrt{n} , Cohen’s algorithm runs in expected linear time. This is precisely

the value of alpha for which the clustering coefficient tends to zero asymptotically, and it is in the range that is relevant for the degree distribution of the World-Wide Web.

- [BePeRo09] J. Bennett, P. Pebay, D. Roe, and D. Thompson. Scalable multi-correlative statistics and principal component analysis with Titan. Technical Report SAND2009-1687, Sandia National Laboratories, March 2009.

Abstract: This report summarizes existing statistical engines in VTK/Titan and presents the recently parallelized multi-correlative and principal component analysis engines. It is a sequel to 'Scalable descriptive and correlative statistics with Titan', which studied the parallel descriptive and correlative engines. The ease of use of these parallel engines is illustrated by the means of C++ code snippets. Furthermore, this report justifies the design of these engines with parallel scalability in mind; then, this theoretical property is verified with test runs that demonstrate optimal parallel speed-up with up to 200 processors.

- [BeThPe09] J. Bennett, D. Thompson, and P. Pebay. Scalable k-means statistics with Titan. Technical Report SAND2009-7855, Sandia National Laboratories, November 2009.

Abstract: This report summarizes existing statistical engines in VTK/Titan and presents both the serial and parallel k-means statistics engines. It is a sequel to earlier reports which studied the parallel descriptive, correlative, multi-correlative, principal component analysis, and contingency engines. The ease of use of the new parallel k-means engine is illustrated by the means of C++ code snippets and algorithm verification is provided. This report justifies the design of the statistics engines with parallel scalability in mind, and provides scalability and speed-up analysis results for the k-means engine.

- [BerryHLP] Jonathan W. Berry, Bruce Hendrickson, Randall A. LaViolette, and Cynthia A. Phillips. Tolerating the community detection resolution limit with edge weighting. *arxiv*, 0903(1072v2), 2009.

Abstract: Communities of vertices within a giant network such as the World-Wide-Web are likely to be vastly smaller than the network itself. However, Fortunato and Barthelemy have proved that modularity maximization algorithms for community detection may fail to resolve communities with fewer than radical $L/2$ edges, where L is the number of edges in the entire network. This resolution limit leads modularity maximization algorithms to have notoriously poor accuracy on many real networks. Fortunato and Barthelemy's argument can be extended to networks with weighted edges as well, and we derive this corollary argument. We conclude that weighted modularity algorithms may fail to resolve communities with fewer than radical $W\epsilon/2$ total edge weight, where W is the

total edge weight in the network and epsilon is the maximum weight of an inter-community edge. If epsilon is small, then small communities can be resolved. Given a weighted or unweighted network, we describe how to derive new edge weights in order to achieve a low epsilon, we modify the 'CNM' community detection algorithm to maximize weighted modularity, and show that the resulting algorithm has greatly improved accuracy. In experiments with an emerging community standard benchmark, we find that our simple CNM variant is competitive with the most accurate community detection methods yet proposed.

- [CaHa09] Alessio Carosui, William Hart, Vitus Leung, and Cynthia Phillips. Problem-specific customization of (integer) linear programming solvers with automatic symbol integration. SAND Report SAND2009-0672, Sandia National Laboratories, 2009.

Abstract: We describe the Solver Utility for Customization with Automatic Symbol Access (SUCASA), a mechanism for generating (integer) linear programming solvers derived from PICO that integrate algebraic modeling constructs. SUCASA allows application developers to access parameters, constraints, and variables from the application algebraic model within PICO. This allows rapid development of problem-specific incumbent heuristics and cutting planes. We briefly describe SUCASA and illustrate its use in two applications: generating graphs with specific degree sequence and scheduling the movements of mobile data collection units in wireless sensor networks.

- [CeBaIb08] A. Cedilnik, J. Baumes, L. Ibanez, S. Megason, and B. Wylie. Integration of information and volume visualization for analysis of cell lineage and gene expression during embryogenesis. In *Proceedings of Visualization and Data Analysis*, volume 6809, January 2008.

Abstract: Dramatic technological advances in the field of genomics have made it possible to sequence the complete genomes of many different organisms. With this overwhelming amount of data at hand, biologists are now confronted with the challenge of understanding the function of the many different elements of the genome. One of the best places to start gaining insight on the mechanisms by which the genome controls an organism is the study of embryogenesis. There are multiple and inter-related layers of information that must be established in order to understand how the genome controls the formation of an organism. One is cell lineage which describes how patterns of cell division give rise to different parts of an organism. Another is gene expression which describes when and where different genes are turned on. Both of these data types can now be acquired using fluorescent laser-scanning (confocal or 2-photon) microscopy of embryos tagged with fluorescent proteins to generate 3D movies of developing embryos. However, analyzing the wealth of resulting images requires tools capable of interactively visualizing several different types of

information as well as being scalable to terabytes of data. This paper describes how the combination of existing large data volume visualization and the new Titan information visualization framework of the Visualization Toolkit (VTK) can be applied to the problem of studying the cell lineage of an organism. In particular, by linking the visualization of spatial and temporal gene expression data with novel ways of visualizing cell lineage data, users can study how the genome regulates different aspects of embryonic development.

- [ChBa09] Peter Chew, Brett Bader, and Alla Rozovskaya. Using DEDICOM for completely unsupervised part-of-speech tagging. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, pages 54–62. Association for Computational Linguistics, 2009.

Abstract: A standard and widespread approach to part-of-speech tagging is based on Hidden Markov Models (HMMs). An alternative approach, pioneered by Schtze (1993), induces parts of speech from scratch using singular value decomposition (SVD). We introduce DEDICOM as an alternative to SVD for part-of-speech induction. DEDICOM retains the advantages of SVD in that it is completely unsupervised: no prior knowledge is required to induce either the tagset or the associations of types with tags. However, unlike SVD, it is also fully compatible with the HMM framework, in that it can be used to estimate emission and transition-probability matrices which can then be used as the input for an HMM. We apply the DEDICOM method to the CONLL corpus (CONLL 2000) and compare the output of DEDICOM to the part-of-speech tags given in the corpus, and find that the correlation (almost 0.5) is quite high. Using DEDICOM, we also estimate part-of-speech ambiguity for each type, and find that these estimates correlate highly with part-of-speech ambiguity as measured in the original corpus (around 0.88). Finally, we show how the output of DEDICOM can be evaluated and compared against the more familiar output of supervised HMM-based tagging.

- [CoGLOr10] R Colbaugh, K Glass, and P Ormerod. Predictability of 'unpredictable' cultural markets. In *Proc 105TH Annual Meeting of the American Sociological Association*, Atlanta, GA, August 2010.

Abstract: In social, economic and cultural situations in which the decisions of individuals are influenced directly by the decisions of others, there is an inherently high level of ex ante unpredictability. We examine the extent to which the existence of social influence may, paradoxically, increase the extent to which the choice which eventually emerges as the most popular (the 'winner') can be identified at a very early stage in the process. Once the process of choice has begun, only a very small number of decisions may be necessary to give a reasonable prospect of being able

to identify the eventual 'winner'. We illustrate this by an analysis of the music download experiments of Salganik et.al. (2006). We derive a practical rule for early identification of the eventual 'winner'. We validate the rule by applying it to similar data not used in the process of constructing the rule.

[CoGIASA10] R Colbaugh and K Glass. Predictive analysis for social diffusion: The role of network communities. In *Proc 105TH Annual Meeting of the American Sociological Association*, Atlanta, GA, August 2010.

Abstract: The diffusion of information and behaviors over social networks is of considerable interest in research fields ranging from sociology to computer science and application domains such as marketing, finance, human health, and national security. Of particular interest is the possibility to develop predictive capabilities for social diffusion, for instance enabling early identification of diffusion processes which are likely to become "viral" and propagate to a significant fraction of the population. Recently we have shown, using theoretical analysis, that the dynamics of social diffusion may depend crucially upon the interactions of social network communities, that is, densely connected groupings of individuals which have only relatively few links between groups. This paper presents an empirical investigation of two related hypotheses which follow from this finding: 1.) inter-community interaction is predictive of the reach of social diffusion and 2.) dispersion of the diffusion phenomenon across network communities is a useful early indicator that the propagation will be "successful". We explore these hypotheses with case studies involving the emergence of the Swedish Social Democratic Party at the turn of the 20th century, the spread of the SARS virus in 2002-2003, and blogging dynamics associated with real world protest activity. These empirical studies demonstrate that network community-based diffusion metrics do indeed possess predictive power, and in fact can be more useful than standard measures.

[CoGIISI09] R Colbaugh and K Glass. Predictive analysis for network dynamics: Part 1 - multi-scale hybrid system modeling. In *Proc IEEE Multi-Conference on Systems and Control*, St Petersburg, Russia, July 2009.

Abstract: This two-part paper presents a new approach to predictive analysis for social processes. In Part I, we begin by identifying a class of social processes which are simultaneously important in applications and difficult to predict using existing methods. It is shown that these processes can be modeled within a multi-scale, stochastic hybrid system framework that is sociologically sensible, expressive, illuminating, and amenable to formal analysis. Among other advantages, the proposed modeling framework enables proper characterization of the interplay between the intrinsic aspects

of a social process (e.g., the “appeal” of a political movement) and the social dynamics which are its realization; this characterization is key to successful social process prediction. The utility of the modeling methodology is illustrated through a case study involving the global SARS epidemic of 2002-2003. Part II of the paper then leverages this modeling framework to develop a rigorous, computationally tractable approach to social process predictive analysis.

[CoGIISI10] R Colbaugh and K Glass. Early warning analysis for social diffusion events. In *Proc IEEE International conference on Intelligence and Security Informatics*, Vancouver, Canada, May 2010. Nominated for Conference Best Paper Award.

Abstract: There is considerable interest in developing predictive capabilities for social diffusion processes, for instance enabling early identification of contentious “triggering” incidents that are likely to grow into large, self-sustaining mobilization events. Recently we have shown, using theoretical analysis, that the dynamics of social diffusion may depend crucially upon the interactions of social network communities, that is, densely connected groupings of individuals which have only relatively few links to other groups. This paper presents an empirical investigation of two hypotheses which follow from this finding: 1.) the presence of even just a few inter-community links can make diffusion activity in one community a significant predictor of activity in otherwise disparate communities and 2.) very early dispersion of a diffusion process across network communities is a reliable early indicator that the diffusion will ultimately involve a substantial number of individuals. We explore these hypotheses with case studies involving emergence of the Swedish Social Democratic Party at the turn of the 20th century, the spread of SARS in 2002-2003, and blogging dynamics associated with potentially incendiary real world occurrences. These empirical studies demonstrate that network community-based diffusion metrics do indeed possess predictive power, and in fact can be significantly more predictive than standard measures.

[CoGIISRD] R Colbaugh and K Glass. Vulnerability analysis of infrastructure networks using aggressive abstractions. Submitted for publication: *Journal of Intelligence Community Research and Development*.

Abstract: Large, complex networks are ubiquitous in nature and society, and there is great interest in developing rigorous, scalable methods for identifying and characterizing their vulnerabilities. This paper presents an approach for analyzing the dynamics of complex networks in which the network of interest is first abstracted to a much simpler, but mathematically equivalent, representation, the required analysis is performed on the abstraction, and analytic conclusions are then mapped back to the original network and interpreted there. We begin by identifying a broad and important class of complex networks which admit vulnerability-preserving,

finite state abstractions, and develop efficient algorithms for computing these abstractions. We then propose a vulnerability analysis methodology which combines these finite state abstractions with formal analytics from theoretical computer science to yield a comprehensive vulnerability analysis process for networks of realworld scale and complexity. The potential of the proposed approach is illustrated with a case study involving a realistic electric power grid model and also with brief discussions of biological and social network examples.

- [CoGIJMS] R Colbaugh and K Glass. Modeling and analysis of social network diffusion: The stochastic hybrid systems approach. *Journal of Mathematical Sociology*. Invited paper.
- [CoGIMTNS10] R Colbaugh and K Glass. Analysis of complex networks using finite state abstraction. In *Proc 19th International Symposium on Mathematical Theory of Networks and Systems*, Budapest, Hungary, July 2010.
- [CoGIPC10] R Colbaugh and K Glass. Influence operations and social networks: What do they say? Presentation at the Phoenix Challenge, April 2010.
- [CoGla10] R Colbaugh and K Glass. Automatically identifying the sources of large internet events. In *Proc IEEE International conference on Intelligence and Security Informatics*, Vancouver, Canada, May 2010.

Abstract: The Internet occasionally experiences large disruptions, arising from both natural and manmade disturbances, and it is of significant interest to develop methods for locating within the network the source of a given disruption (i.e., the network element(s) whose perturbation initiated the event). This paper presents a near real-time approach to realizing this logical localization objective. The proposed methodology consists of three steps: 1.) data acquisition/preprocessing, in which publicly available measurements of Internet activity are acquired, “cleaned”, and assembled into a format suitable for computational analysis, 2.) event characterization via tensor factorization-based time series analysis, and 3.) localization of the source of the disruption through graph theoretic analysis. This procedure provides a principled, automated approach to identifying the root causes of network disruptions at “whole-Internet” scale. The considerable potential of the proposed analytic method is illustrated through a computer simulation study and empirical analysis of a recent, large-scale Internet disruption.

- [CoGlb10] R Colbaugh and K Glass. Estimating sentiment orientation in social media for intelligence monitoring and analysis. In *Proc IEEE International conference on Intelligence and Security Informatics*, Vancouver, Canada, May 2010.

Abstract: This paper presents a computational approach to inferring the sentiment orientation of “social media” content (e.g., blog posts) which focuses on the challenges associated with Web-based analysis. The proposed methodology formulates the task as one of text classification, models the data as a bipartite graph of documents and words, and uses this framework to develop a semi-supervised sentiment classifier that is well-suited for social media domains. In particular, the proposed algorithm is capable of combining prior knowledge regarding the sentiment orientation of a few documents and words with information present in unlabeled data, which is abundant online. We demonstrate the utility of the approach by showing it outperforms several standard methods for the task of inferring the sentiment of online movie reviews, and illustrate its potential for security informatics through a case study involving the estimation of Indonesian public sentiment regarding the July 2009 Jakarta hotel bombings.

- [CoGoG110] R Colbaugh, J Gosler, and K Glass. Some intelligence analysis problems and their graph formulations. *Journal of Intelligence Community Research and Development*, (315), 2010.

Abstract: This paper considers three important classes of intelligence analysis questions, shows how these questions can be naturally formulated as graph problems, and demonstrates that analysis based on this formulation yields insights and understanding which would be difficult to obtain using other methods. In order to make explicit the relevance and practical utility of graph methods for intelligence questions, we organize our discussion around specific real world problems of current interest in the counterproliferation and counterterrorism domains.

- [DoLA09] Courtney Dornburg, Laura Matzen, Travis Bauer, and Laura McNamara. Working memory load as a novel tool for evaluating visual analytics. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 217–218. IEEE, 2009.

Abstract: The current visual analytics literature highlights design and evaluation processes that are highly variable and situation dependent, which raises at least two broad challenges. First, lack of a standardized evaluation criterion leads to costly re-designs for each task and specific user community. Second, this inadequacy in criterion validation raises significant uncertainty regarding visualization outputs and their related decisions, which may be especially troubling in high consequence environments like those of the Intelligence Community. As an attempt to standardize the “apples and oranges” of the extant situation, we propose the creation of standardized evaluation tools using general principles of human cognition. Theoretically, visual analytics enables the user to see information in a way that should attenuate the user’s memory load and increase

the user's task-available cognitive resources. By using general cognitive abilities like available working memory resources as our dependent measures, we propose to develop standardized evaluative capabilities that can be generalized across contexts, tasks, and user communities.

- [LaGlCo09] R Lavolette, K Glass, and R Colbaugh. Deep information from limited observation of robust yet fragile systems. *Physica A*, 388:3283–3287, October 2009.

Abstract: We show how one can completely reconstruct even moderately optimized configurations of the Forest Fire model with relatively few observations. We discuss the relationship between the deep information from limited observations (DILO) to the robust-yet-fragile (RYF) property of the Forest Fire model and propose that DILO may be a general property of RYF complex systems.

- [LeLa09] Vitus J. Leung and Randall A. LaViolette. Sensitivity of the performance of a simple exchange model to its topology. SAND Report SAND2009-3380, Sandia National Laboratories, 2009.

Abstract: We study a simple exchange model in which price is fixed and the amount of a good transferred between actors depends only on the actors' respective budgets and the existence of a link between transacting actors. The model induces a simply-connected but possibly multi-component bipartite graph. A trading session on a fixed graph consists of a sequence of exchanges between connected buyers and sellers until no more exchanges are possible. We deem a trading session “feasible” if all of the buyers satisfy their respective demands. If all trading sessions are feasible the graph is deemed “successful”, otherwise the feasibility of a trading session depends on the order of the sequence of exchanges. We demonstrate that topology is important for the success of trading sessions on graphs. In particular, for the case that supply equals demand for each component of the graph, we prove that the graph is successful if and only if the graph consists of components each of which are complete bipartite. For the case that supply exceeds demand, we prove that the other topologies also can be made successful but with finite reserve (i.e., excess supply) requirements that may grow proportional to the number of buyers. Finally, with computations for a small instance of the model, we provide an example of the wide range of performance in which only the connectivity varies. These results taken together place limits on the improvements in performance that can be expected from proposals to increase the connectivity of sparse exchange networks.

- [MaMc10] Laura Matzen, Laura McNamara, Kerstan Cole, Alisa Bandlow, Courtney Dornburg, and Travis Bauer. Proposed Working Memory Measures for Evaluating Information Visualization Tools. In *BELIV'10: BEyond time and errors: novel evaluation methods for Information Visualization*, 2010.

Abstract: The current information visualization literature highlights design and evaluation processes that are highly variable and situation dependent, which raises at least two broad challenges. First, lack of a standardized evaluation criterion leads to costly re-designs for each task and specific user community. Second, this inadequacy in criterion validation raises significant uncertainty regarding visualization outputs and their related decisions, which may be especially troubling in high consequence environments like those of intelligence analysts. We seek ways to standardize the “apples and oranges” of the extant situation through tools based upon general principles of human cognition. Theoretically, information visualization tools enable the user to see information in a way that should attenuate the user’s memory load and increase the user’s task-available cognitive resources. By using general cognitive abilities, like available working memory resources, as a dependent measure, we propose standardized evaluative capabilities can be generalized across contexts, tasks, and user communities

[MaMcEuro10] Laura Matzen, Laura McNamara, Alisa Bandlow, Kerstan Cole, and Courtney Dornburg. Evaluating information visualization tools with measures of working memory load. In *Eurographics/IEEE Symposium on Visualization*, 2010.

[MuShTe09] R.C. Murphy, S. Shinde, and J. Teifel. Network grand challenge custom architecture task final report. Technical Report SAND2009-7589, Sandia National Laboratories, September 2009.

[NGC] The Networks Grand Challenge.

URL ngc.sandia.gov

[OIBaChCCIM10] R.A. Oldfield, B.W Bader, and P. Chew. Applying high-performance computing to multilingual document clustering. Submitted to CCIM Computational Science Research Highlights, 2010.

[OIBaChSiam10] R.A. Oldfield, B.W Bader, and P. Chew. Supporting multilingual document clustering on the Cray XT3. In *SIAM Conference on Parallel Processing and Scientific Computing*, February 2010.

Abstract: This talk describes a parallel multilingual documentclustering application designed for the Cray XT3 (Red-Storm) system at Sandia National Laboratories. Our application is unique among HPC applications because it provides interactive visualization and analysis capability by spanning three different architectures: the Cray XT3, a visualization cluster, and Netezza Data Warehouse appliance. We will discuss design, scalability challenges, and results for several large multilingual data sets, including the Bible, the Quran, and proceedings of the European Parliament.

[OIWiDa09] R.A. Oldfield, A. Wilson, G. Davidson, and C. Ulmer. Access to external resources using service-node proxies. In *Proceedings of the Cray User Group Meeting*, Atlanta, GA, May 2009. SAND 2009-2773 C.

[Old10] R.A. Oldfield. Hey! you got your DWA in my HPC : Experiences integrating Netezza and Cray XT3. Invited talk, Cray Hybrid Solutions Summit, June 2010. SAND 2010-4676 P.

[PeTh08] P. Pebay and D. Thompson. Scalable descriptive and correlative statistics with Titan. Technical Report SAND2008-8260, Sandia National Laboratories, December 2008.

Abstract: This report summarizes the existing statistical engines in VTK/Titan and presents the parallel versions thereof which have already been implemented. The ease of use of these parallel engines is illustrated by the means of C++ code snippets. Furthermore, this report justifies the design of these engines with parallel scalability in mind; then, this theoretical property is verified with test runs that demonstrate optimal parallel speed-up with up to 200 processors.

[PeTh09] P. Pebay and D. Thompson. Parallel contingency statistics with titan. Technical Report SAND2009-6006, Sandia National Laboratories, September 2009.

Abstract: This report summarizes existing statistical engines in VTK/Titan and presents the recently parallelized contingency statistics engine. It is a sequel to previous work which studied the parallel descriptive, correlative, multi-correlative, and principal component analysis engines. The ease of use of this new parallel engines is illustrated by the means of C++ code snippets. Furthermore, this report justifies the design of these engines with parallel scalability in mind; however, the very nature of contingency tables prevent this new engine from exhibiting optimal parallel speed-up as the aforementioned engines do. This report therefore discusses the design trade-offs we made and study performance with up to 200 processors.

[Ro10] David G. Robinson. Statistical language analysis for automatic exfiltration event detection. In *Proceedings of the 2010 Workshop on Quantitative Methods in Defense and National Security*, Statistics in Defense and National Security. American Statistical Association, May 2010.

Abstract: This paper discusses the recent development a statistical approach for the automatic identification of anomalous network activity that is characteristic of exfiltration events. This approach is based on the language processing method referred to as latent dirichlet allocation (LDA). Cyber security experts currently depend heavily on a rule-based framework for initial detection of suspect network events. The application of the rule set typically results in an extensive list of suspect network events that

are then further explored manually for suspicious activity. The ability to identify anomalous network events is heavily dependent on the experience of the security personnel wading through the network log. Limitations of this approach are clear: rule-based systems only apply to exfiltration behavior that has previously been observed, and experienced cyber security personnel are rare commodities. Since the new methodology is not a discrete rule-based approach, it is more difficult for an insider to disguise the exfiltration events. A further benefit is that the methodology provides a risk-based approach that can be implemented in a continuous, dynamic or evolutionary fashion. This permits suspect network activity to be identified early with a quantifiable risk associated with decision making when responding to suspicious activity.

[Ro10NIPS] David G. Robinson. Natural language processing for exfiltration event detection. In *Proceedings of the 2010 Neural Information Processing Systems Conference*, December 2010.

[SeBaKo10] Mark Sears, Brett Bader, and Tamara Kolda. Algorithms for tensors and tensor models. Sand report, Sandia National Laboratories, 2010.

[ShTi09-P] Tim Shead. NGC integration and HPC analysis.

URL ngc.sandia.gov/assets/documents/titan-eab10_SAND2010-5300P.ppt

[ShTi10-P] Tim Shead. NGC/Titan impact.

URL ngc.sandia.gov/assets/documents/titan-eab09_SAND2009-6318P.ppt

[StHF10] William Stubblefield. Interaction design assessment of P2. 2010.

[Titan] The Titan informatics toolkit.

URL titan.sandia.gov

[VAST10] IEEE VAST 2010 challenge.

URL hcil.cs.umd.edu/localphp/hcil/vast10/index.php

[WiCh10] Andrew T. Wilson and Peter A. Chew. Term weighting schemes for latent dirichlet allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 465–473, Los Angeles, California, June 2010. Association for Computational Linguistics.

URL www.aclweb.org/anthology/N10-1070

Abstract: Many implementations of Latent Dirichlet Allocation (LDA), including those described in Blei et al. (2003), rely at some point on the removal of stopwords, words which are assumed to contribute little to the

meaning of the text. This step is considered necessary because otherwise high-frequency words tend to end up scattered across many of the latent topics without much rhyme or reason. We show, however, that the “problem” of high-frequency words can be dealt with more elegantly, and in a way that to our knowledge has not been considered in LDA, through the use of appropriate weighting schemes comparable to those sometimes used in Latent Semantic Indexing (LSI). Our proposed weighting methods not only make theoretical sense, but can also be shown to improve precision significantly on a non-trivial cross-language retrieval task.

- [WyBa09] Brian Wylie and Jeff Baumes. A unified toolkit for information and scientific visualization. In *Proceedings of Visualization and Data Analysis*, volume 7243, January 2009.

Abstract: We present an expansion of the popular open source Visualization Toolkit (VTK) to support the ingestion, processing, and display of informatics data. The result is a flexible, component-based pipeline framework for the integration and deployment of algorithms in the scientific and informatics fields. This project, code named “Titan”, is one of the first efforts to address the unification of information and scientific visualization in a systematic fashion. The result includes a wide range of informatics-oriented functionality: database access, graph algorithms, graph layouts, views, charts, UI components and more. Further, the data distribution, parallel processing and client/server capabilities of VTK provide an excellent platform for scalable analysis.

- [WyBaSh08] Brian Wylie, Jeffrey Baumes, Timothy Shead, and John Greenfield. Information visualization with VTK. Tutorial at VisWeek 2008, July 2008.

Abstract: The merging of information visualization and scientific visualization is an active area of research, development and discussion. Recently the Visualization Toolkit (VTK) has been expanded with new data structures and a large set of filters, views, database adapters and other components that support the processing of informatics data. Our tutorial gives a detailed overview of the new components and demonstrates their use within the flexible pipeline architecture. Attendees will learn the basics of using VTK for information visualization with a focus on building quick, functional applications. The tutorial will also demonstrate how users can extend the current capabilities to include their own functionality.

- [WyBaSh08a] Brian Wylie, Jeffrey Baumes, and Timothy Shead. Gspace: A linear time graph layout. Presented at 2008, July 2008.

Abstract: We describe G-Space (Geodesic Space), a straightforward linear time layout algorithm that draws undirected graphs based purely on their topological features. The algorithm is divided into two phases. The

first phase is an embedding of the graph into a 2-D plane using geodesic distances as coordinates. These coordinates are computed with the same process used by HDE (High-Dimensional Embedding) algorithms. In our case we do a Low-Dimensional Embedding (LDE), and directly map the geodesic distances into a two dimensional geometric space. The second phase is the resolution of the many-to-one mappings that frequently occur within the low dimensional embedding. The resulting layout appears to have advantages over existing methods: it can be computed rapidly, and it can be used to answer topological questions quickly and intuitively.

DISTRIBUTION:

- 1 MS 0359 D. Chavez, LDRD Office, 1911
- 1 MS 0899 Technical Library, 8944 (electronic copy)





Sandia National Laboratories