Data Mining on Attributed Relationship Graphs (ARGs)

LDRD Day

September 19, 2007

.....

Tamara G. Kolda
Principal Investigator and PMTS
Informatics and Decision Sciences Dept. (Org. 8962)

Team Members: Brett Bader (1416), Bruce Hendrickson (1415), Ann Yoshimura (8116), Joe Kenny (8961), Travis Bauer (6341)

Plus collaborations with: Peter Chew (6341), Danny Dunlavy (1411), Philip Kegelmeyer (8962)

LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories

Attributed Relationship Graphs

- **Graph**
  - *Nodes* represent entities such as people, places, or objects
  - *Edges* represent connections between entities
- **Attributed Relationship Graphs (ARGs)**
  - Nodes and edges have *types*, allowing multiple types in one graph
  - Both nodes and edges can have *attributes* (names, dates, etc.)
- **ARGS are used by intelligence analysts as a way to integrate data from disparate sources**

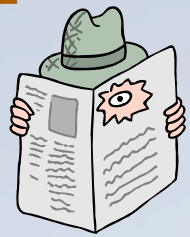
ARGs are also known as *Semantic Graphs*

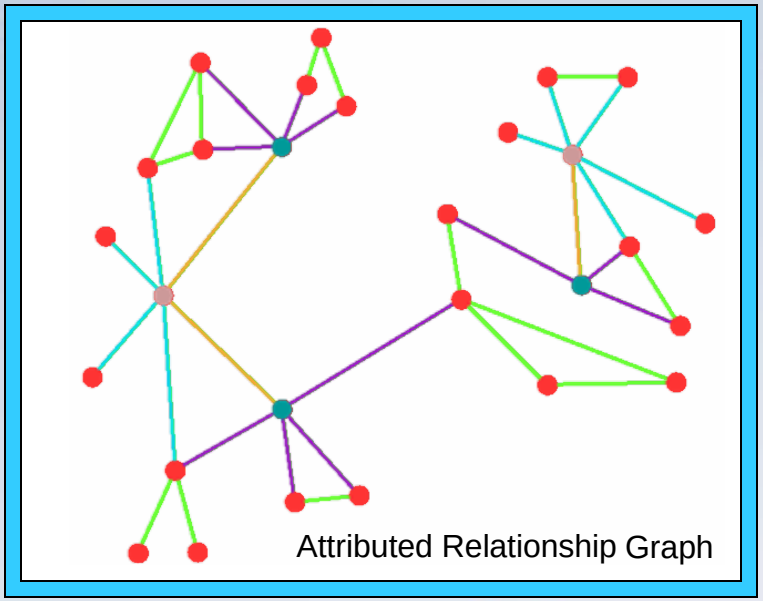

email


telephone

open source

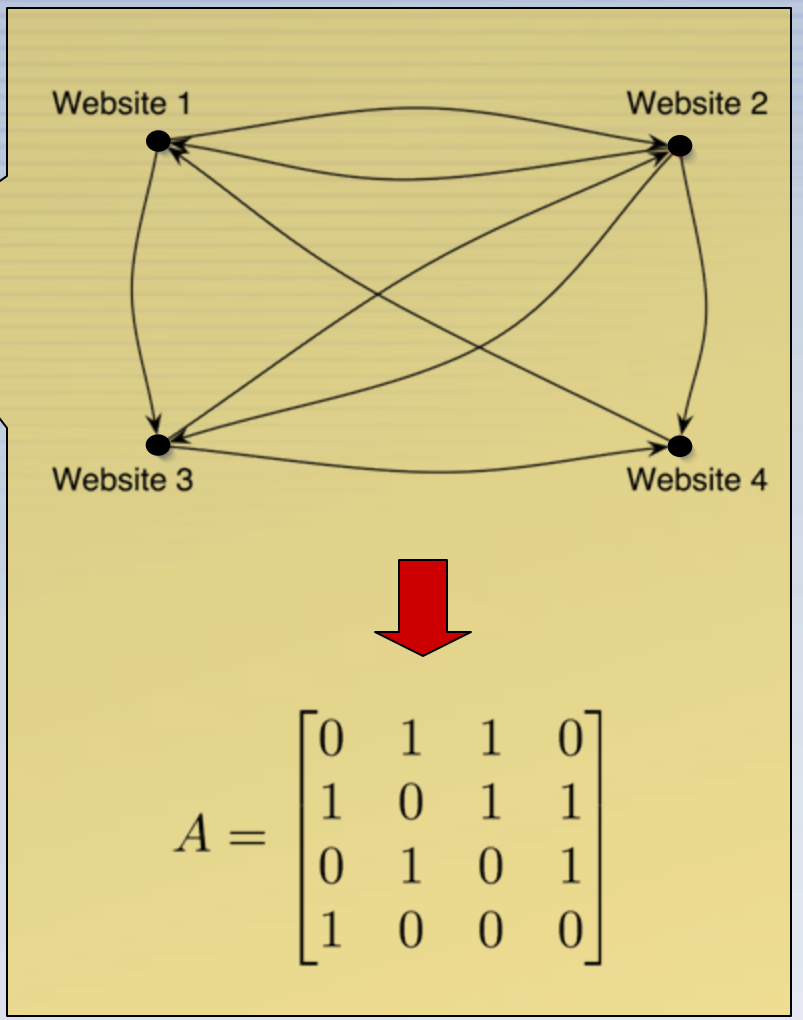

cell phone


intelligence reports



LDRD Purpose: Develop Latent Semantic Analysis for ARGs

- **Matrix decompositions are a tool for graph analysis**
  - Hubs & Authorities (HITS)
    - Matrix stores hyperlinks between web pages
    - Compute singular value decomposition (SVD)
    - Kleinberg (1998)
  - Google PageRank
    - Stochastic matrix stores hyperlinks plus random jumps
    - Compute leading eigenvector
    - Page et al. (1998)
  - Latent Semantic Indexing (LSI)
    - Matrix links documents and terms
    - Compute singular value decomposition (SVD)
    - Dumais et al. (1988)
- **Tensors provide a natural representation for ARGs**
  - Tensor decompositions can be used for similar analyses!



LDRD Approach: Use Tensor Decompositions to Analyze ARGS

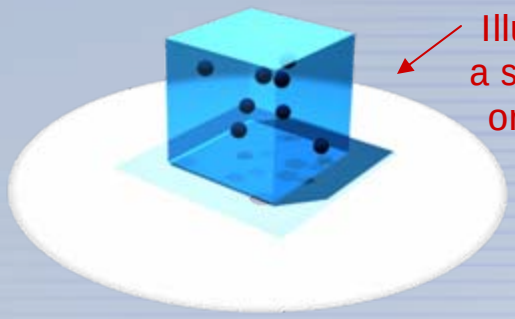
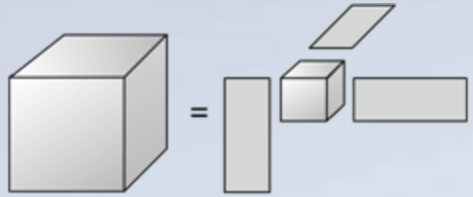


Illustration of a sparse third-order tensor

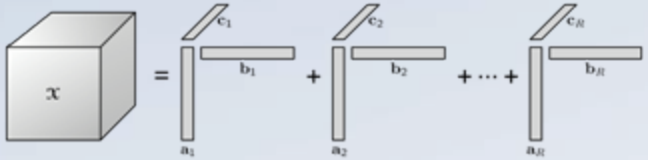
- **A tensor is a multi-way array**
- **Not to be confused with tensor fields such as stress tensors.**
- **We have focused specially on sparse tensors!**

- **Tensor decompositions reveal latent structure in the entries of a tensor**
  - Higher-order analogues of matrix SVD/PCA
  - Date back to Hitchcock (1927)
  - Popularized
    - 1970s in psychometrics
    - 1980s in chemometrics
    - 1990s in signal processing
    - 2000s in data mining, etc.

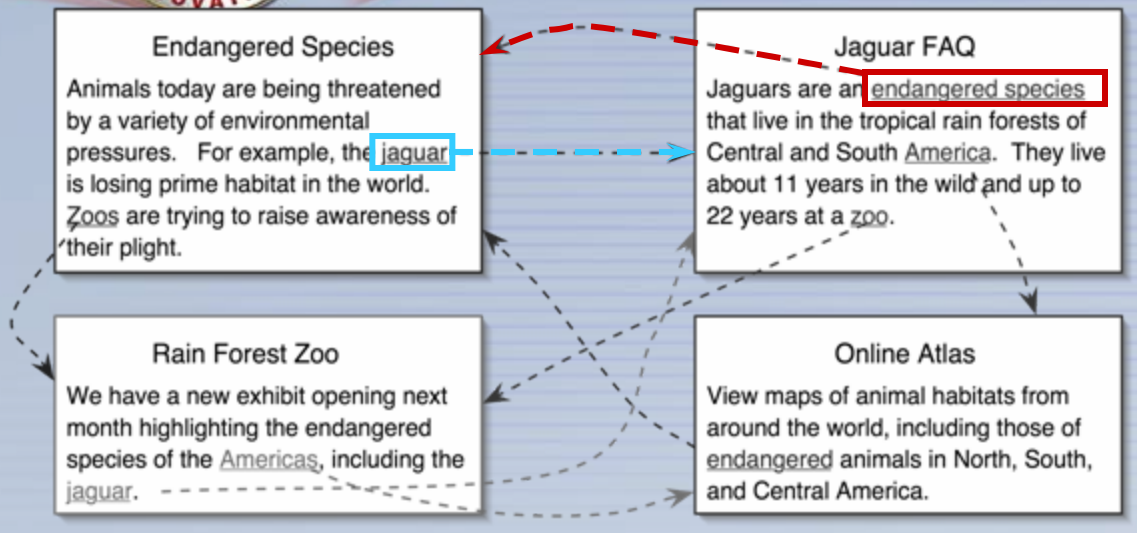
Tucker (1966)



CANDECOMP/ PARAFAC (1970)



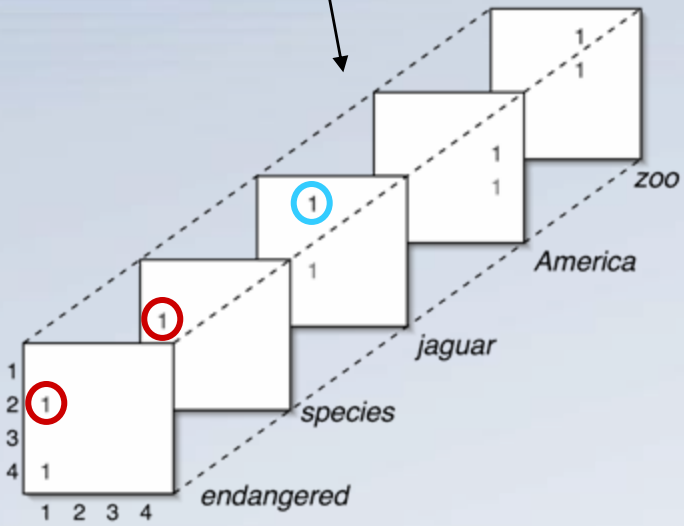
TOPHITS – A Three-Dimensional View of the Web

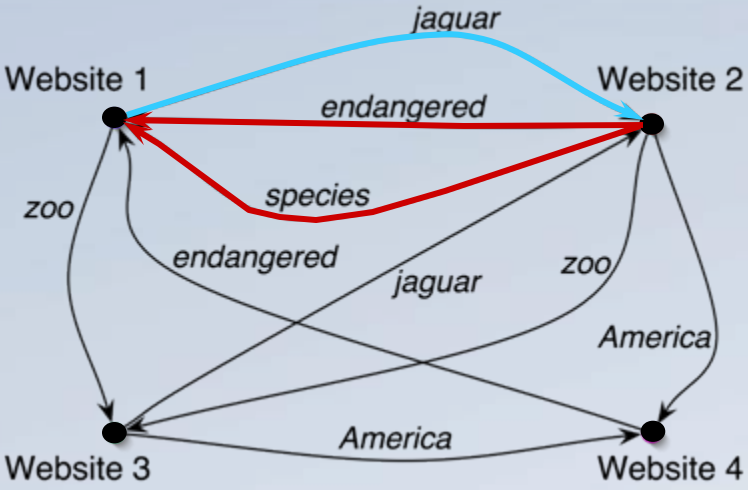


Tensor Definition

$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{page } j \\ & \text{with term } k \\ 0 & \text{otherwise} \end{cases}$$

Observe that this tensor is very sparse!





TOPHITS Terms & Authorities on Sample Data

1st Principal Factor

.23	JAVA	.86	java.sun.com
.18	SUN	.38	developers.sun.com
.17	PLATF		
.16	SOLAR		
.16	DEVEL		
.15	EDITIO		
.15	DOWN		
.14	INFO		
.12	SOFTW		
.12	NO-RE		

2nd Principal Factor

.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACU		
.16	SEAR		
.16	NEWS		
.16	LIBRA		
.16	COMP		
.12	LEHIC		

3rd Principal Factor

.15	NO-READABLE-TEXT	.97	www.ibm.com
.15	IBM		
.12	SERVI		
.12	WEBS		
.12	WEB		
.11	DEVEL		
.11	LINUX		
.11	RESO		
.11	TECHI		
.10	DOWN		

4th Principal Factor

.26	INFORMATION	.87	www.pueblo.gsa.gov
.24	FEDERAL	.24	www.irs.gov
.23	CITIZE		
.22	OTHE		
.19	CENTI		
.19	LANGU		
.15	U.S		
.15	PUBLI		
.14	CONS		
.13	FREE		

6th Principal Factor

.26	PRESIDENT	.87	www.whitehouse.gov
.25	NO-READABLE-TEXT	.18	www.irs.gov
.25	BUSH		
.25	WELC		
.17	WHITE		
.16	U.S		
.15	HOUS		
.13	BUDG		
.13	PRESI		
.11	OFFIC		

12th Principal Factor

.75	OPTIMIZATION	.35	www.palisade.com
.58	SOFTW		
.08	DECIS		
.07	NEOS		
.06	TREE		
.05	GUIDE		
.05	SEARC		
.05	ENGIN		
.05	CONT		
.05	ILOG		

13th Principal Factor

.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACRO		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

16th Principal Factor

.50	WEATHER	.81	www.weather.gov
.24	OFFICE		
.23	CENTE		
.19	NO-REA		
.17	ORGAN		
.15	NWS		
.15	SEVER		
.15	FIRE		
.15	POLICY		
.14	CLIMAT		

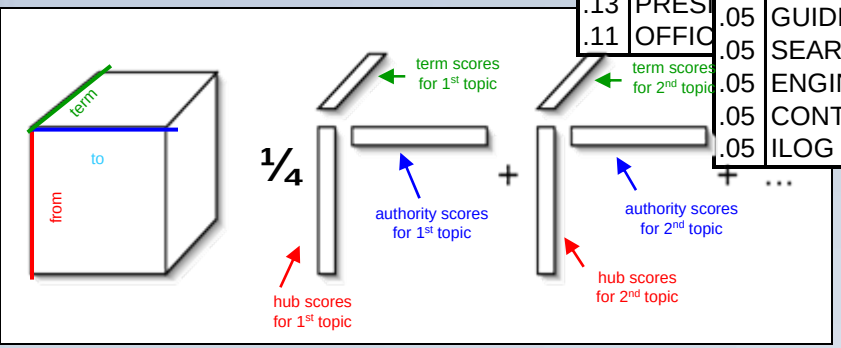
19th Principal Factor

.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

Main Idea: Ability to automatically group and *label* web pages according to their importance and topic.

$$x_{ijk} = \begin{cases} \frac{1}{\log(w_k)+1} & \text{if } i \to j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

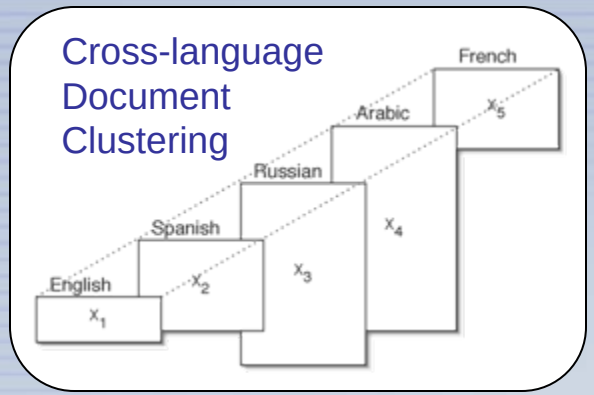
W_k = # unique links using term k

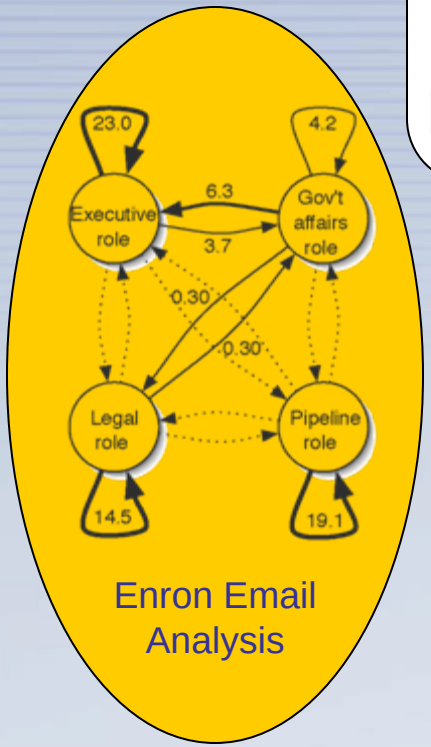


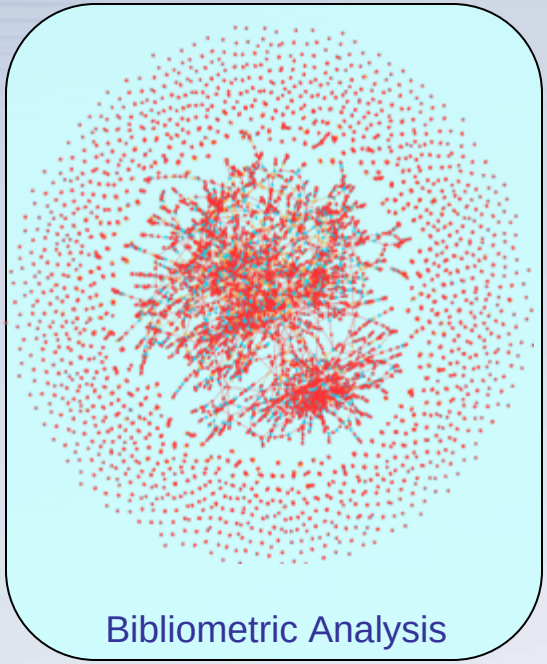
Tensor PARAFAC

Significance: Tensors are a New Approach to Graph & Data Analysis

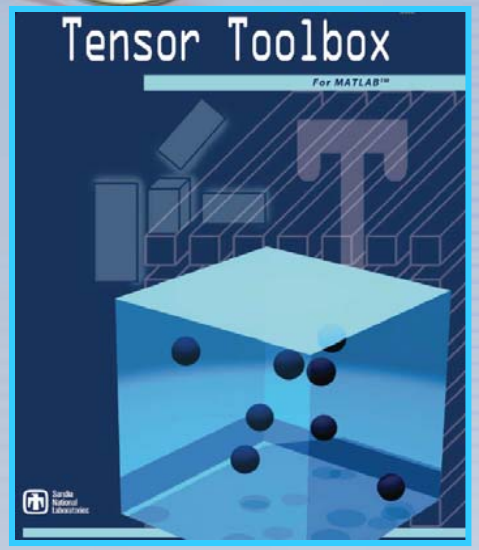
- **TOPHITS for higher-order web link analysis**
  - Automatic grouping and labeling
- **Enron email text analysis**
  - Automatic temporal topic discovery
- **Tutorial: Mining large time-evolving data using matrix and tensor tools**
  - Temporal data analysis, including anomaly detection
- **Cross-language document clustering**
  - Text analysis / information retrieval
- **Enron email temporal pattern analysis**
  - Automatic “role” discovery of entities in the network
- **Bibliometric analysis**
  - Feature generation, entity resolution



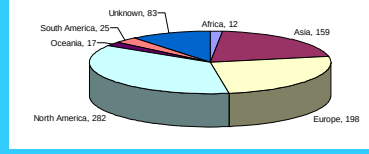




Results Include Software Supporting Data and Graph Analysis

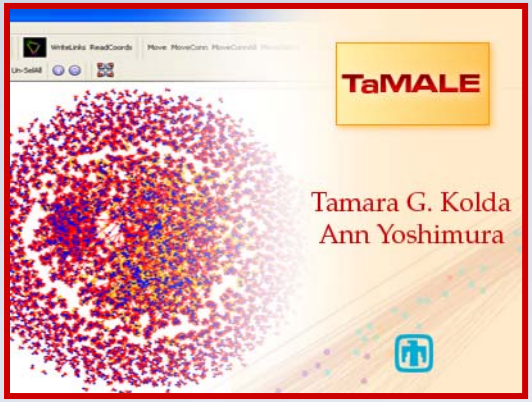


- Tensor Toolbox for MATLAB
- B. W. Bader (1416) & T. G. Kolda (8962)
- Version 2 released externally Sept. 2006; 775+ Downloads from all over the world
- Publications on the toolbox in ACM Transactions on Mathematics Software and SIAM J. Scientific Computing
- Unique capability: Methods for large-scale, sparse and structured tensors



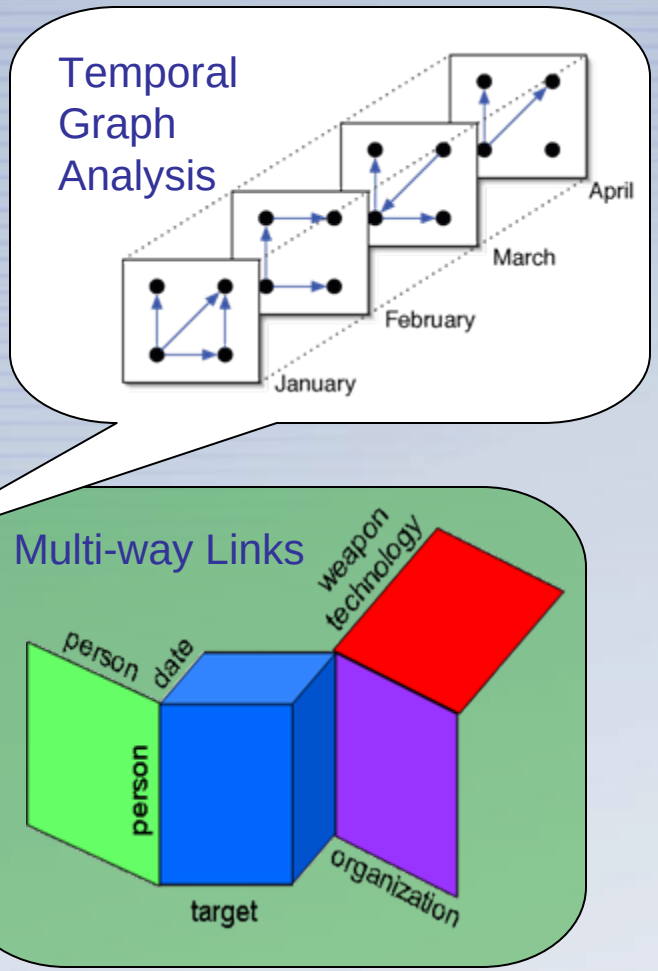
Downloads by Continent

- TaMALE = Tensor Multi-Attribute Link Explorer
- T. G. Kolda (8962) and A. Yoshimura (8116)
- Import, visualization, and analysis of ARGs
- Released internally in March 2007
- Version 2.0 now available



Conclusion: Tensors Provide a Novel Graph Analysis Capability

- **Impact is novel tensor-based data/graph analysis techniques**
  - Includes software supporting new capabilities
- **More graph analysis on the horizon**
  - Enabling All-Threat Analysis Through Intelligent Filtering of Network Traffic
    - DHS LDRD FY07-09
    - PI: Jamie Van Randwyk (8965)
  - Network Discovery, Characterization & Prediction
    - Grand Challenge LDRD FY08-10
    - PI: Bruce Hendrickson (1415)
  - Leveraging Multi-way Linkages on Heterogeneous Data
    - EPS LDRD FY08-10
    - PI: Kolda



For more information, contact:
Tammy Kolda
294-4769, tgkolda@sandia.gov