Data Mining on Attributed Relationship Graphs (ARGs)

Sandia National Laboratories

Tamara G. Kolda (PI), (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)

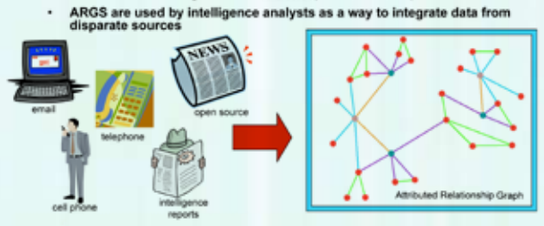
Plus collaborations with: Peter Chew (6341), Danny Dunlavy (1411), Philip Kegelmeyer (8962)

LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Attributed Relationship Graphs

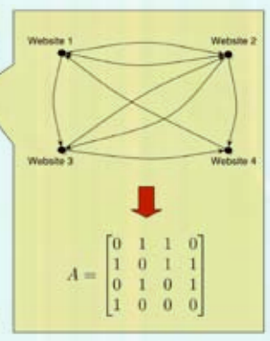
- **Graph**
  - *Nodes* represent entities such as people, places, or objects
  - *Edges* represent connections between entities
- **Attributed Relationship Graphs (ARGs)**
  - Nodes and edges have *types*, allowing multiple types in one graph
  - Both nodes and edges can have *attributes* (names, dates,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

ARGs are also known as *Semantic Graphs*

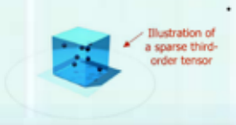


LDRD Purpose: Develop Latent Semantic Analysis for ARGs

- **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!

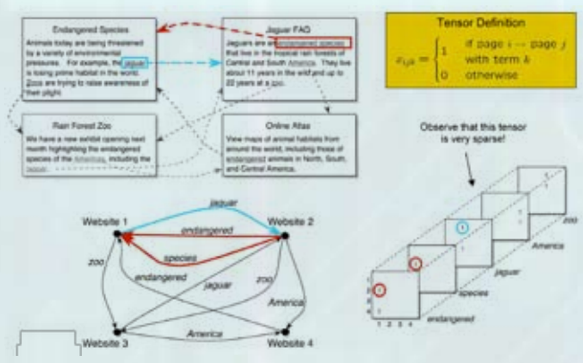


LDRD Approach: Use Tensor Decompositions to Analyze ARGS

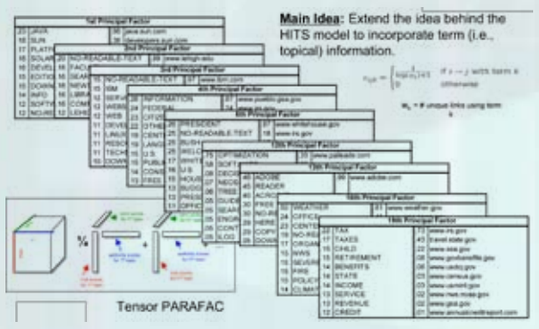


- **Tensor decompositions reveal latent structure in the entries of a tensor**
  - Higher-order analogues of matrix SVD/PCA
  - Date back to Hitchcock (1927)
  - Popularized
    - 1970s in psychometrics
    - 1980s in chemometrics
    - 1990s in signal processing
    - 2000s in data mining, etc.
- **A tensor is a multi-way array**
- **Not to be confused with tensor fields such as stress tensors.**
- **We have focused specially on sparse tensors!**

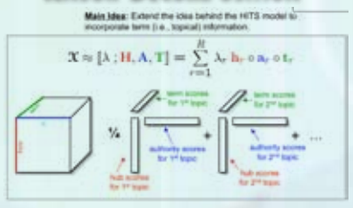
TOPHITS – A Three-Dimensional View of the Web



TOPHITS Terms & Authorities on Sample Data

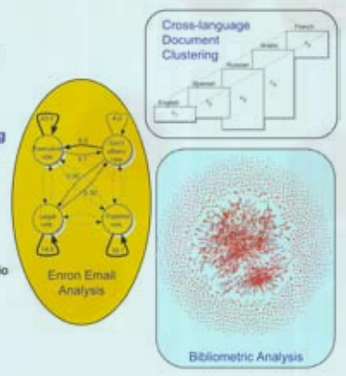


Detail: Topical HITS (TOPHITS) Tensor Decomposition



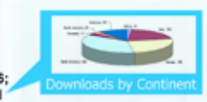
Significance: Tensors are a New Approach to Graph & Data Analysis

- TOPHITS for higher-order web link analysis
- Enron email text analysis
  - With Mike Berry, U. Tennessee
- Tutorial: Mining large time-evolving data using matrix and tensor tools
  - With C. Faloutsos, J. Sun (CMU)
- Cross-language document clustering
- Enron email temporal pattern analysis
  - With R. Harshman, U. Western Ontario
- Bibliometric analysis
- Social network analysis



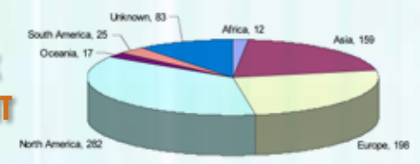
Results: New Algorithms & Software for Data Analysis

- **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**



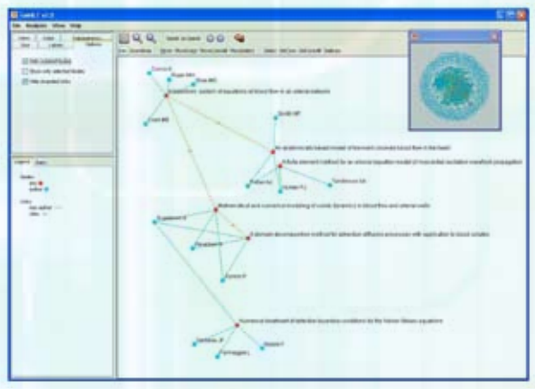
- **TaMALE = Tensor Multi-Attribute Link Explorer**
- **T. G. Kolda (8962) and A. Yoshimura (8116)**
- **Import, visualization, and analysis of ARGs**
- **Released internally in March 2007**
- **Version 2.0 now available**

Detail: Tensor Toolbox Downloads by Continent



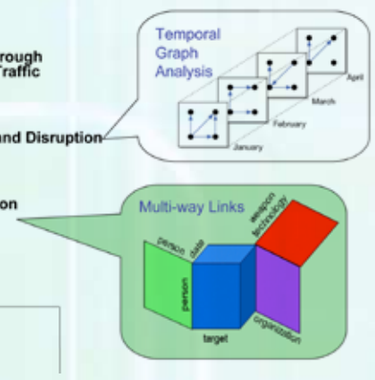
Unique downloads (by IP address) since Sept. 2006

Detail: Bibliometric Graph in TaMALE



This Project is Over, But There is 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, Prediction and Disruption**
  - Grand Challenge LDRD FY08-10
  - PI: Bruce Hendrickson (1415)
- **Leveraging Multi-way Linkages on Heterogeneous Data**
  - EPS LDRD FY08-10
  - PI: Kolda



Detail: Other Papers and Selected Presentations

- **Journal Papers**
  - Brett W. Bader and Tamara G. Kolda. **Efficient MATLAB computations with sparse and factored tensors.** *SIAM Journal on Scientific Computing*, Accepted for publication, July 2007.
  - Brett W. Bader and Tamara G. Kolda. **Algorithm 862: MATLAB tensor classes for fast algorithm prototyping.** *ACM Transactions on Mathematical Software*, 32(4):635–653, December 2006. (doi:10.1145/1186785.1186794)
- **Conference & Workshop Papers**
  - Tamara G. Kolda, Brett W. Bader, and Joseph P. Kenny. **Higher-order web link analysis using multilinear algebra.** In *ICDM 2005: Proceedings of the 5th IEEE International Conference on Data Mining*, pages 242–249, November 2005. (doi:10.1109/ICDM.2005.77)
  - Brett W. Bader, Michael W. Berry, and Murray Browne. **Discussion tracking in Enron email using PARAFAC.** Text Mining Workshop, April 2007.
  - Peter A. Chew, Brett W. Bader, Tamara G. Kolda, and Ahmed Abdelali. **Cross-language information retrieval using PARAFAC2.** In *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 143–152. ACM Press, 2007. (doi:10.1145/1281192.1281211)
  - Brett W. Bader, Richard Harshman, and Tamara G. Kolda. **Temporal analysis of semantic graphs using ASALSAN.** In *ICDM 2007: Proceedings of the 7th IEEE International Conference on Data Mining*, November 2007. To appear.
- **Submitted Journal Papers**
  - Tamara G. Kolda and Brett W. Bader. **Tensor decompositions and applications.** Submitted to *SIAM Review*, August 2007.
- **Other papers**
  - Teresa M. Selee, Tamara G. Kolda, W. Philip Kegelmeyer, and Joshua D. Griffin. **Extracting clusters from large datasets with multiple similarity measures using IMSCAND.** To appear in CSRI2007 proceedings, August 2007.
  - Daniel M. Dunlavy, Tamara G. Kolda, and W. Philip Kegelmeyer. **Multilinear algebra for analyzing data with multiple linkages.** Technical Report SAND2006-2079, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, April 2006. Revised version to appear in *Array Based Graph Algorithms*
  - Tamara G. Kolda. **Multilinear operators for higher-order decompositions.** Technical Report SAND2006-2081, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, April 2006.
- **Refereed Tutorial**
  - Christos Faloutsos, Tamara G. Kolda, and Jimeng Sun, **Mining Large Time-evolving Data Using Matrix and Tensor Tools.** Tutorial at SDM07, SIGMOD07, ICML07, KDD07.

Tamara G. Kolda (PI), 925-294-4769 or tgkolda@sandia.gov

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2007-5945C CreativeArts 98429.LW.8.07