Data Mining on Attributed Relationship Graphs (ARGS)

SAND2007-5944C

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**

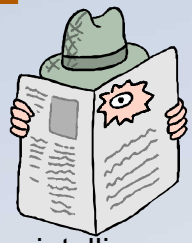
ARGs are also known as *Semantic Graphs*

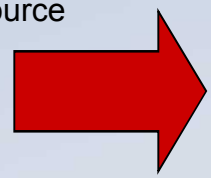
email

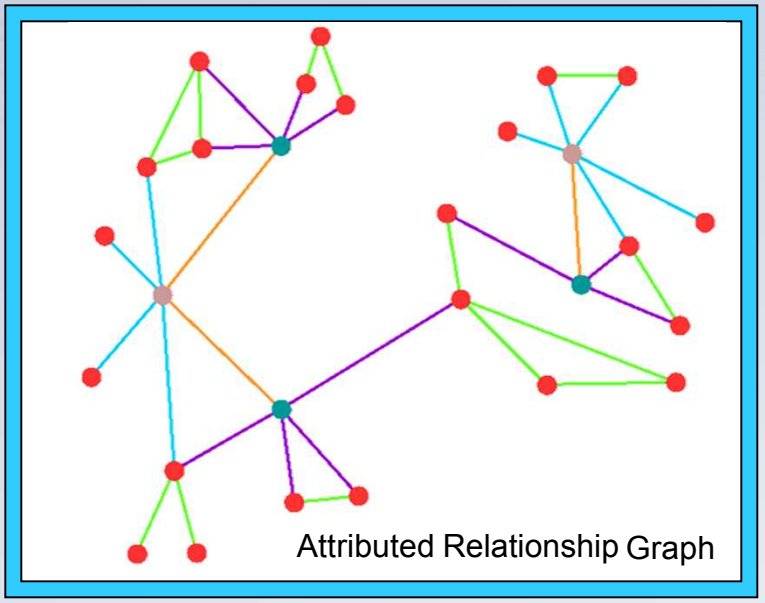
telephone

open source

cell phone

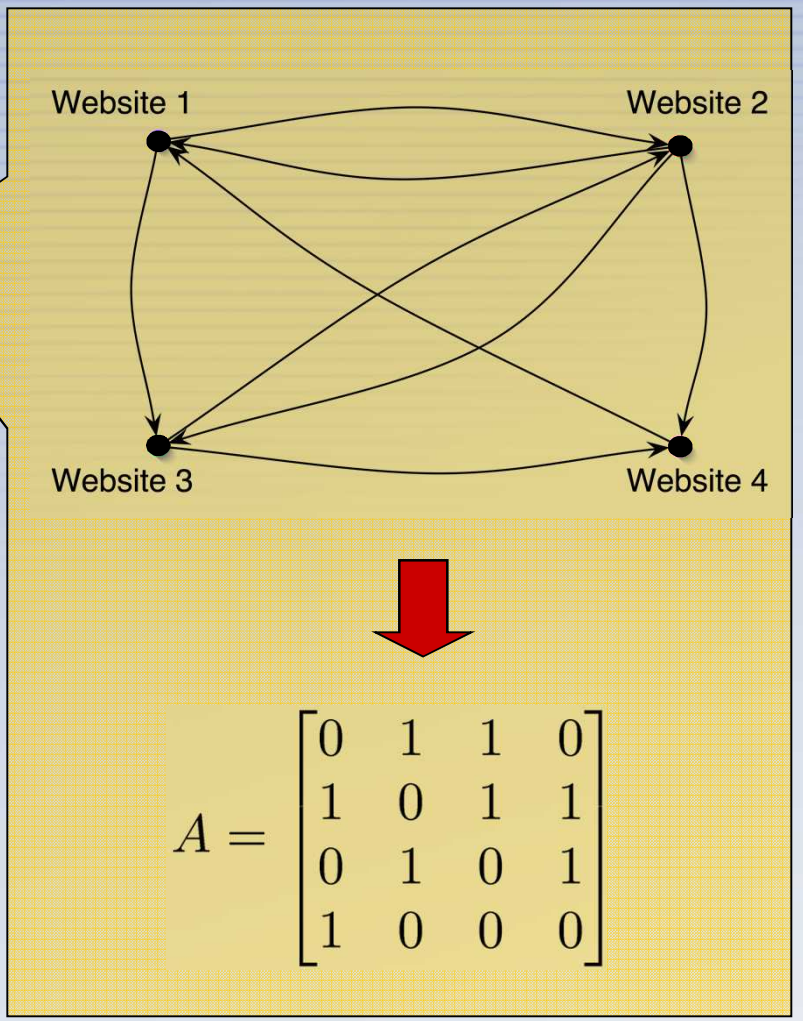

intelligence reports





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

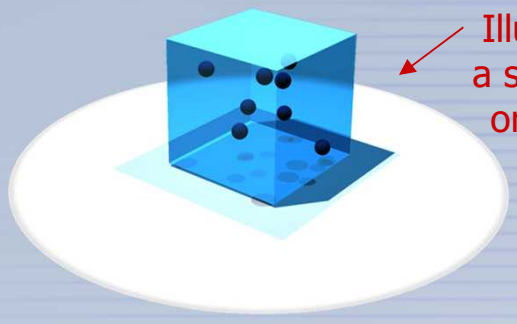
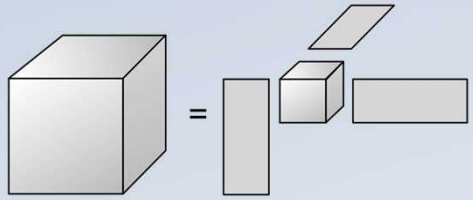


Illustration of a sparse third-order tensor

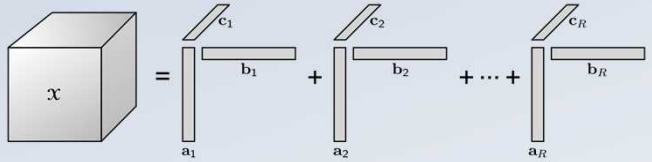
- **A tensor is a multi-way array**
- **Not to be confused with tensor fields such as stress tensors.**
- **We have focused specially on sparse tensors!**

- **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.

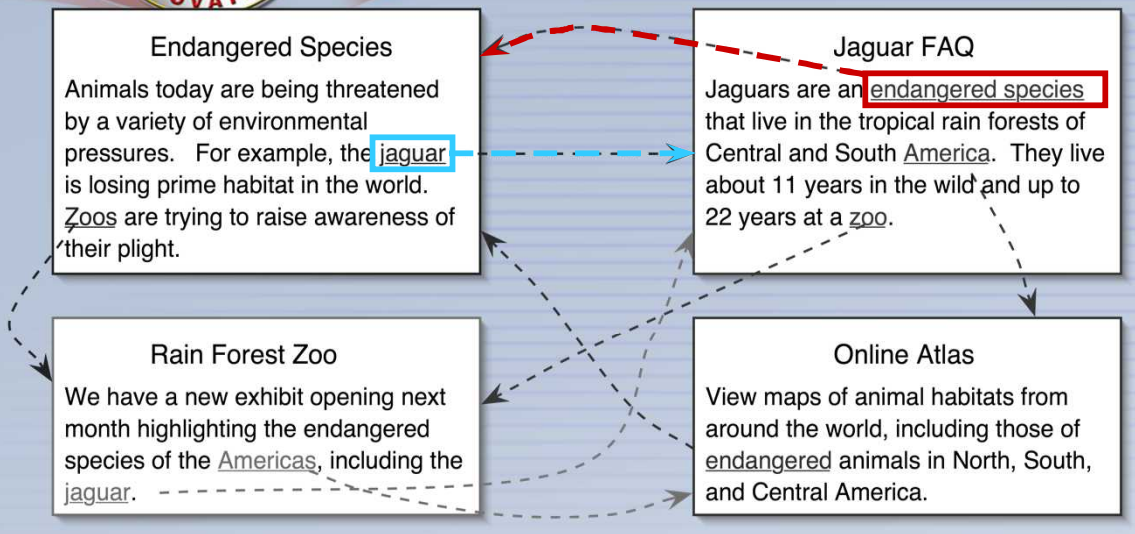
Tucker (1966)



CANDECOMP/ PARAFAC (1970)



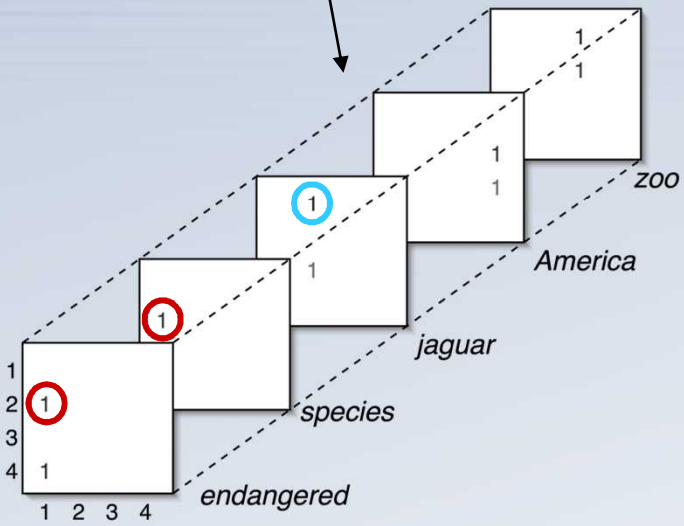
TOPHITS – A Three-Dimensional View of the Web

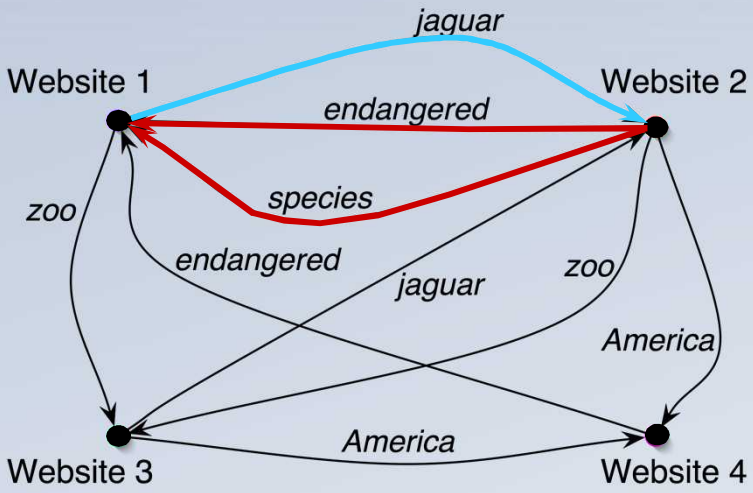


Tensor Definition

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

Observe that this tensor is very sparse!





LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT
Sandia National Laboratories

TOPHITS Terms & Authorities on Sample Data

Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

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

w_k = # unique links using term k

1st Principal Factor

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

2nd Principal Factor

.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACU		
.16	SEAR		
.16	NEWS		
.16	LIBRA		
.16	COMF		
.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	LANG		
.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

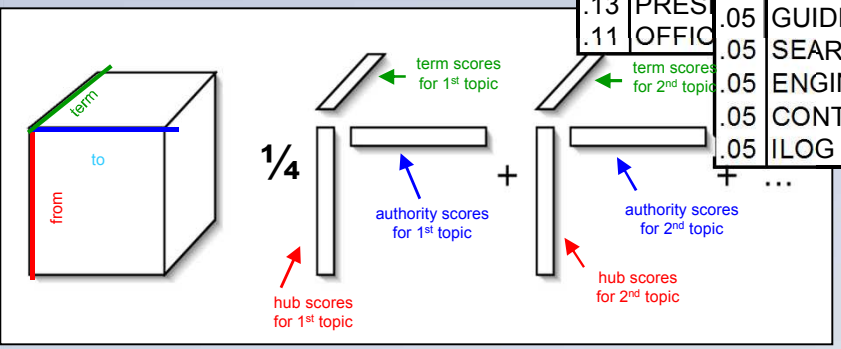
.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACRO		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

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

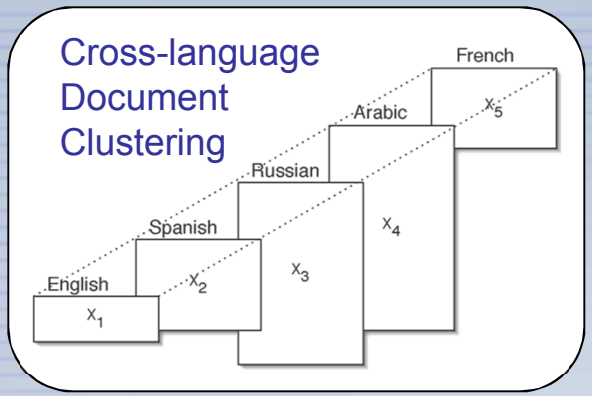
.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

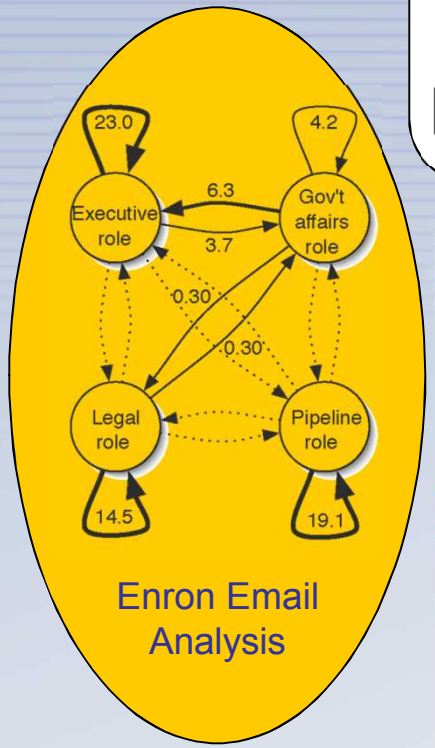


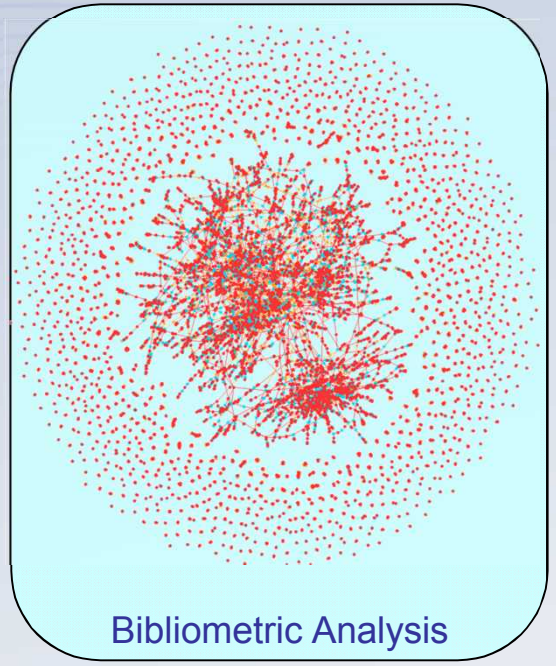
Tensor PARAFAC

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

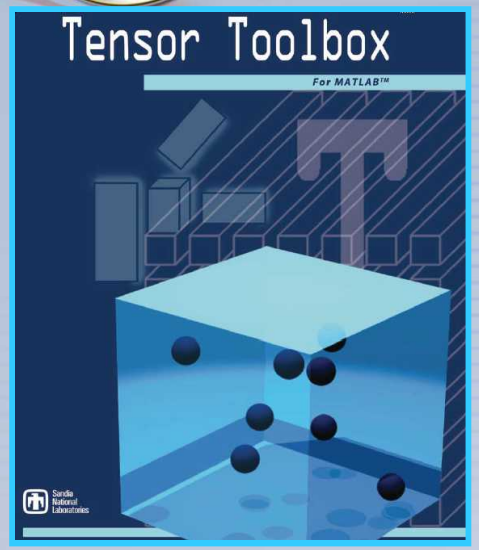
- **TOPHITS for higher-order web link analysis**
- **Enron email text analysis**
  - With Mike Berry, U. Tennessee
- **Tutorial: Mining large time-evolving data using matrix and tensor tools**
  - With C. Faloutsos, J. Sun (CMU)
- **Cross-language document clustering**
- **Enron email temporal pattern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



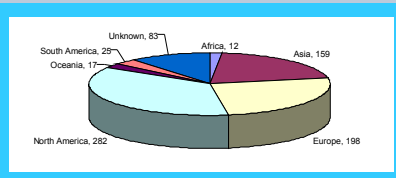




Results: New Algorithms & Software for Data Analysis

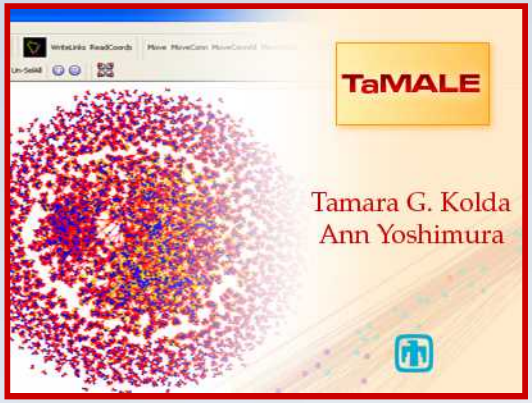


- Tensor Toolbox for MATLAB
- B. W. Bader (1416) & T. G. Kolda (8962)
- Version 2 released externally Sept. 2006; 775+ Downloads from all over the world
- Publications on the toolbox in ACM Transactions on Mathematics Software and SIAM J. Scientific Computing
- Unique capability: Methods for large-scale,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



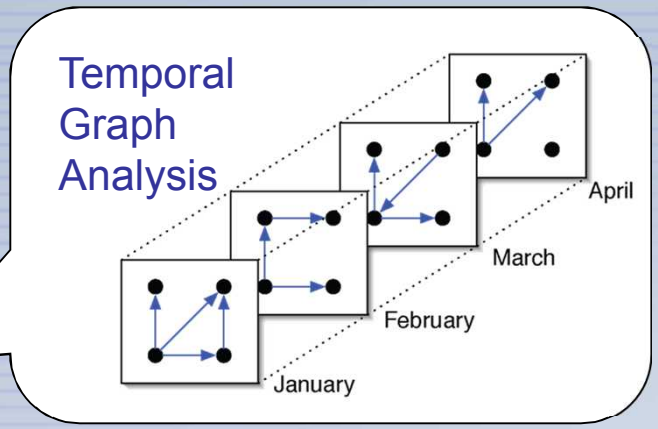
Downloads by Continent

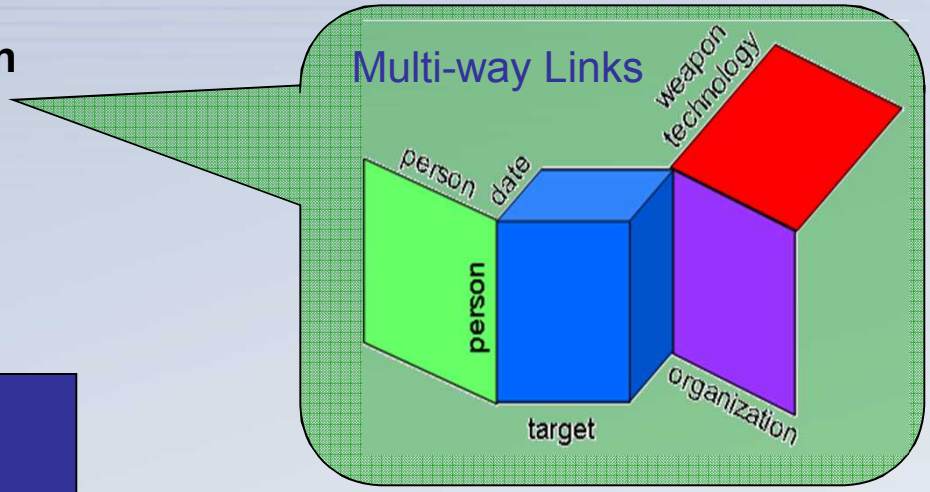
- TaMALE = Tensor Multi-Attribute Link Explorer
- T. G. Kolda (8962) and A. Yoshimura (8116)
- Import, visualization, and analysis of ARGs
- Released internally in March 2007
- Version 2.0 now available



This project is over, but there is more graph analysis on the horizon

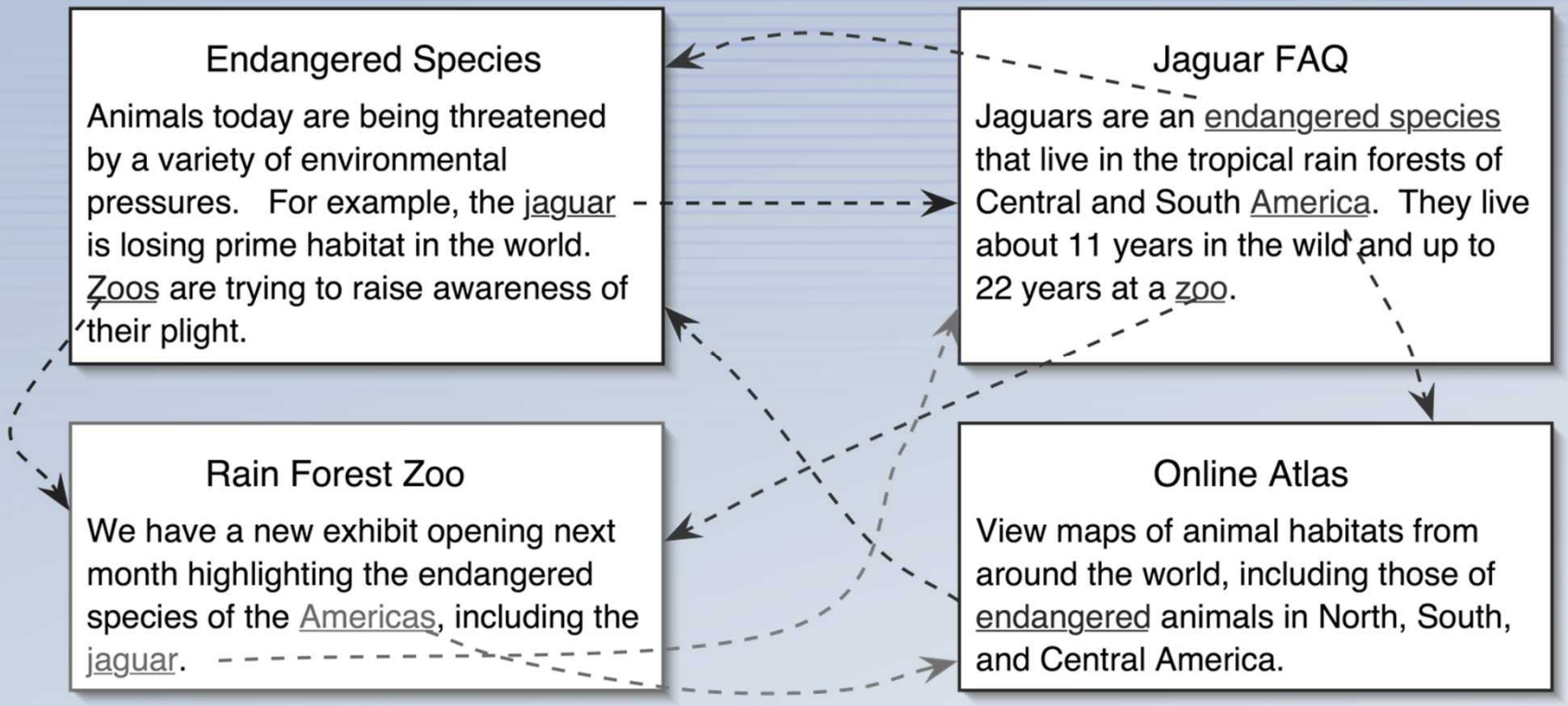
- **Enabling All-Threat Analysis Through Intelligent Filtering of Network Traffic**
  - DHS LDRD FY07-09
  - PI: Jamie Van Randwyk (8965)
- **Network Discovery, Prediction and Disruption**
  - Grand Challenge LDRD FY08-10
  - PI: Bruce Hendrickson (1415)
- **Leveraging Multi-way Linkages on Heterogeneous Data**
  - EPS LDRD FY08-10
  - PI: Kolda





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

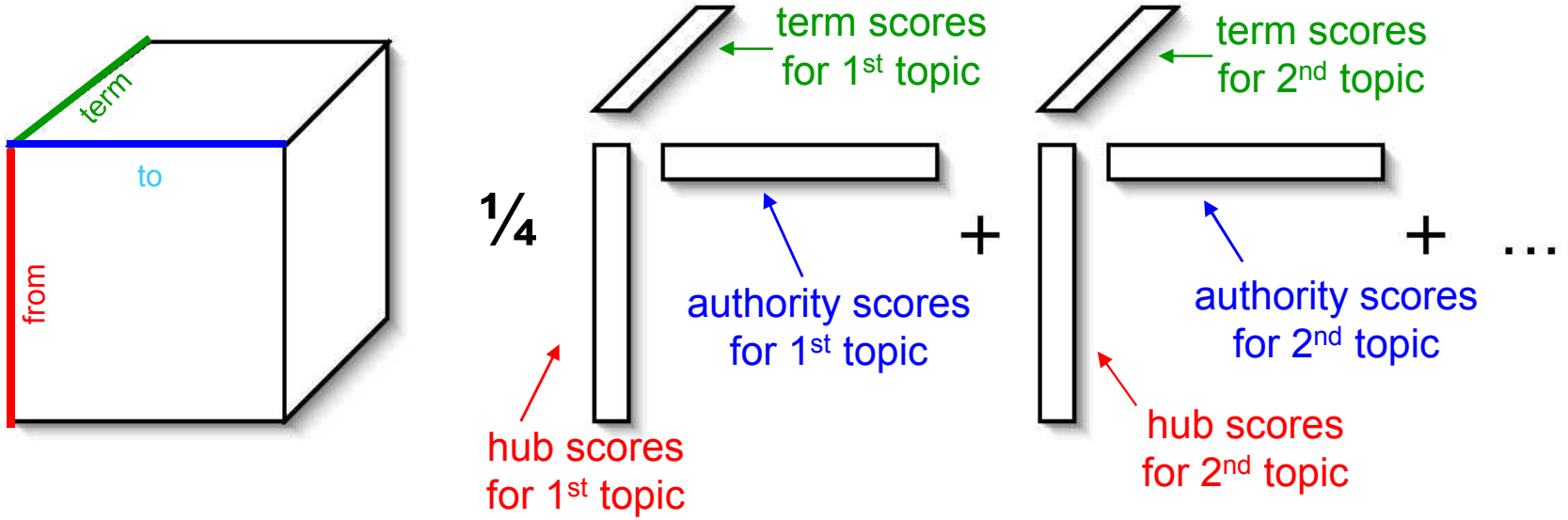
Detail: Example Web Pages



DETAIL: Topical HITS (TOPHITS) Tensor Decomposition

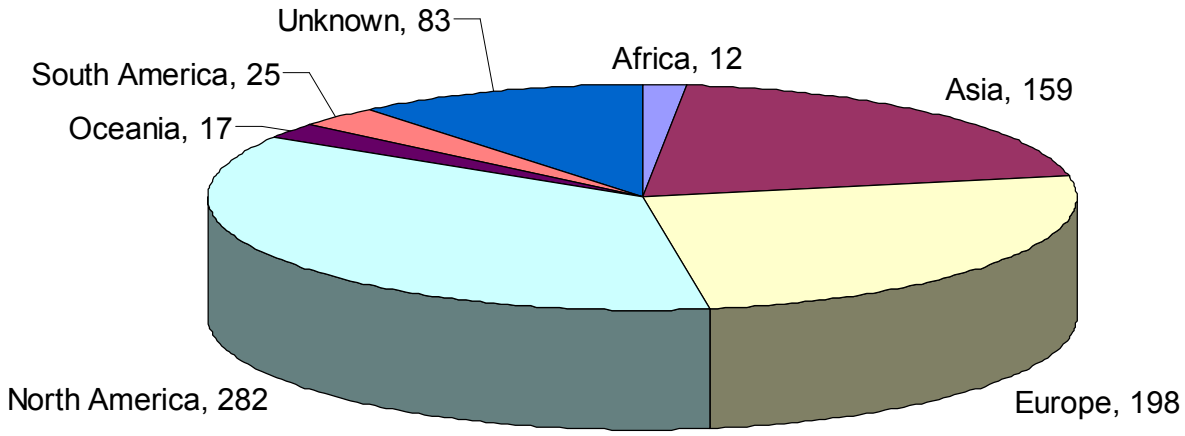
Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

$$\mathcal{X} \approx [\![\lambda\,; \mathbf{H}, \mathbf{A}, \mathbf{T}]\!] = \sum_{r=1}^{R} \lambda_r\, \mathbf{h}_r \circ \mathbf{a}_r \circ \mathbf{t}_r$$



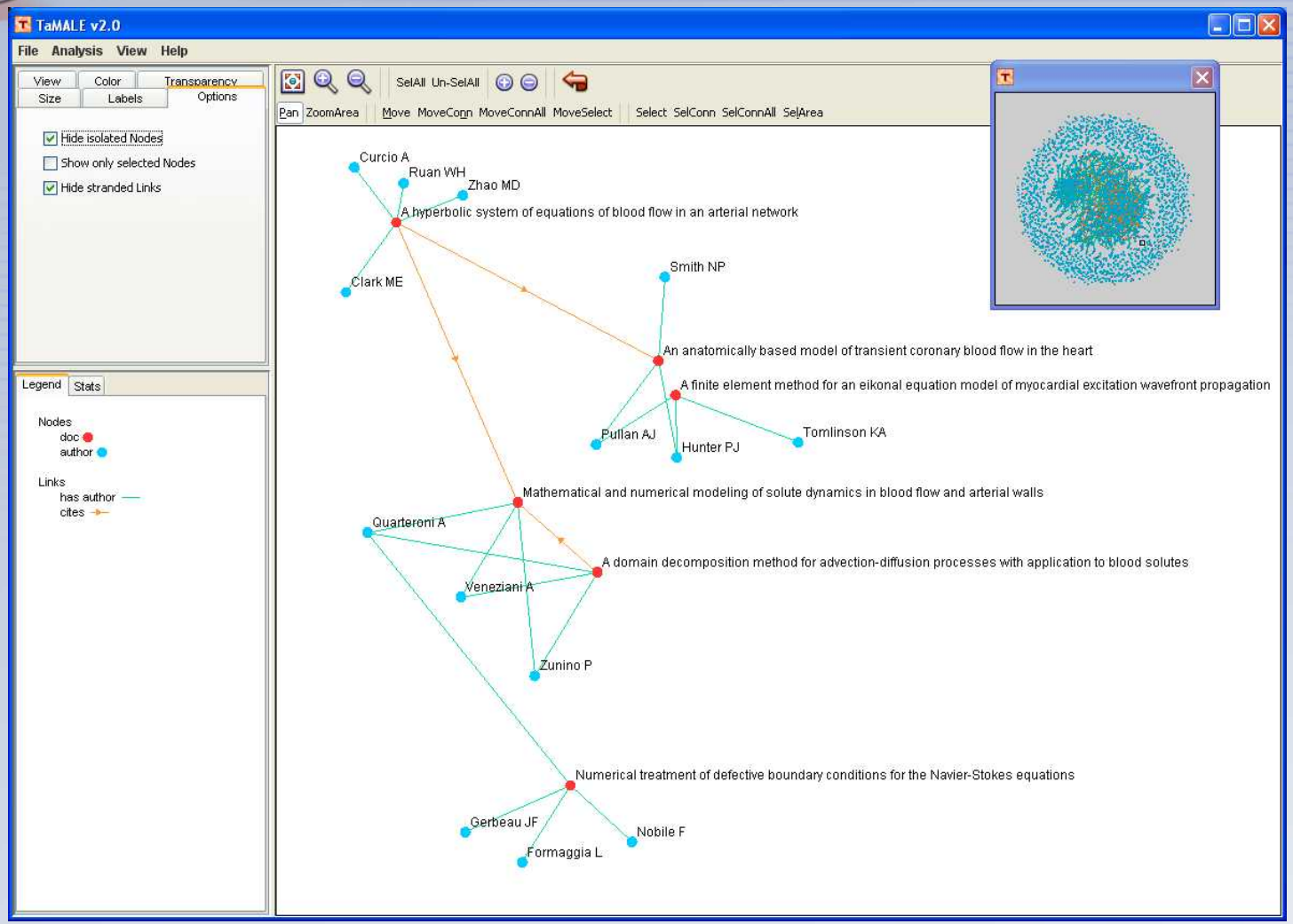
T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In ICDM 2005: Proceedings of the 5th IEEE International Conference on Data Mining, pages 242–249, November 2005. (doi:10.1109/ICDM.2005.77)

Detail: Tensor Toolbox Downloads by Continent



Unique downloads (by IP address) since Sept. 2006

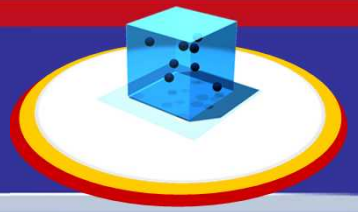
Detail: Bibliometric Graph in TaMALE



Detail: Refereed Papers

- **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.

Detail: Other Papers and Presentations



- **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.