SAND2010-4090P

Algebraic and Tensor Methods for Anomaly Detection

Brett Bader
Senior Member of Technical Staff
Computer Science and Informatics

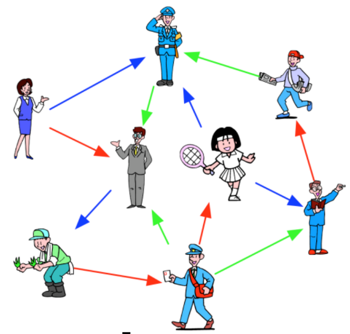
June 17, 2010

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.

Robust data analysis requires appropriate data abstractions and algorithms

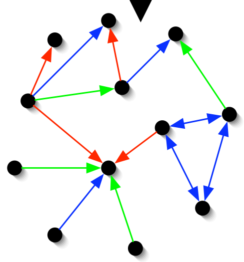
Sandia uses *semantic graphs* and *tensors* as unifying data abstractions

- Supports rich relationship-centered analysis
- Combines large, heterogeneous data corpora
- Different abstractions support different analytics

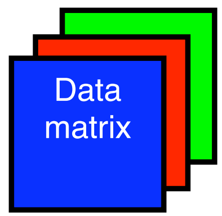


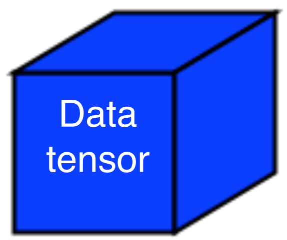
Real world networks
- Social networks
- Cyber traffic
- etc.

Raw, unstructured text
- Newspaper articles
- Web documents
- etc.



Semantic Graph



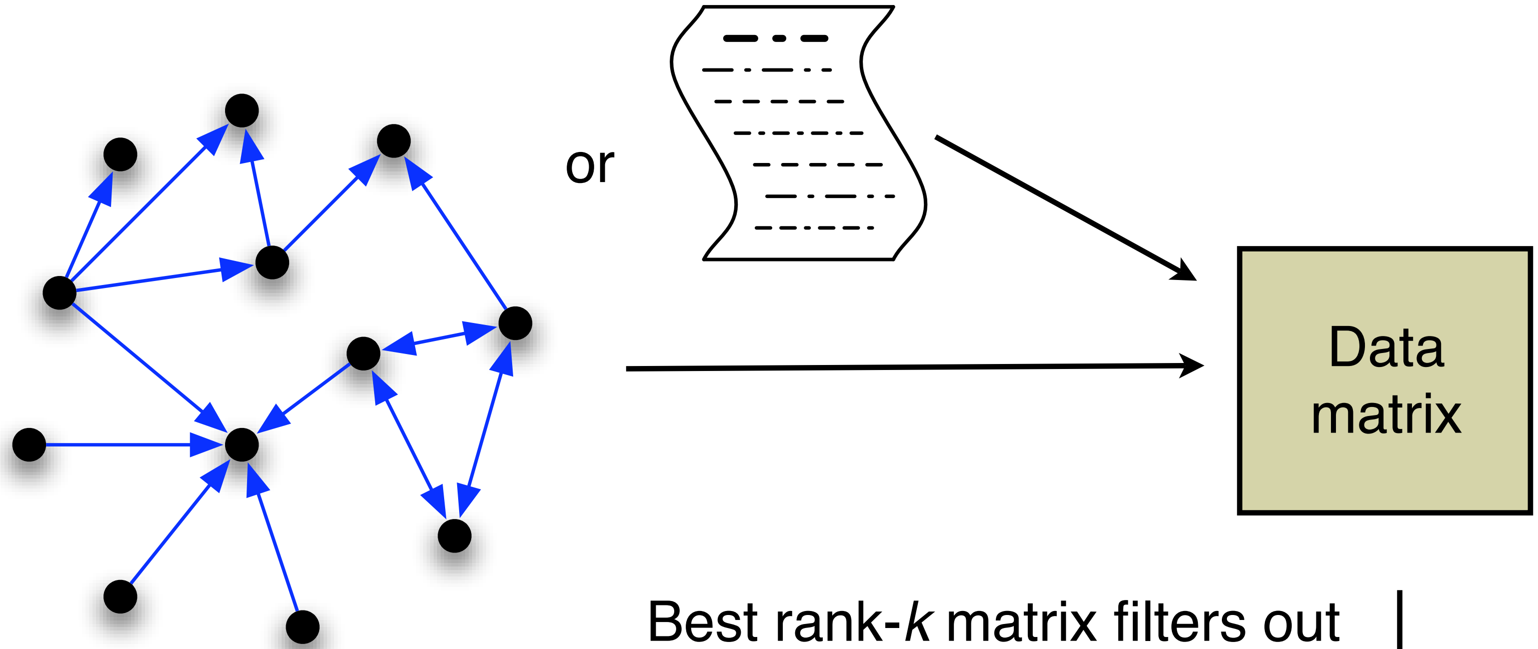


Matrices & Tensors

- Graph algorithms, discrete math
- Short paths, connection subgraphs, subgraph isomorphism

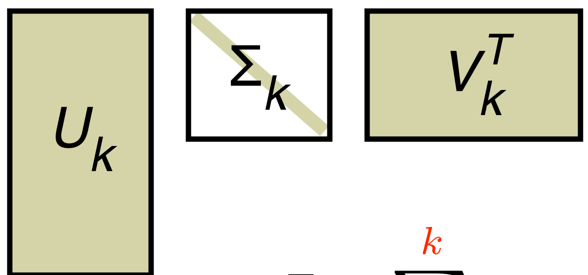
- Linear and multilinear algebra, statistics/probability
- Ranking, clustering

Traditional Analysis



Best rank-k matrix filters out noise and captures "latent" information, which improves certain data mining tasks

Truncated SVD



$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

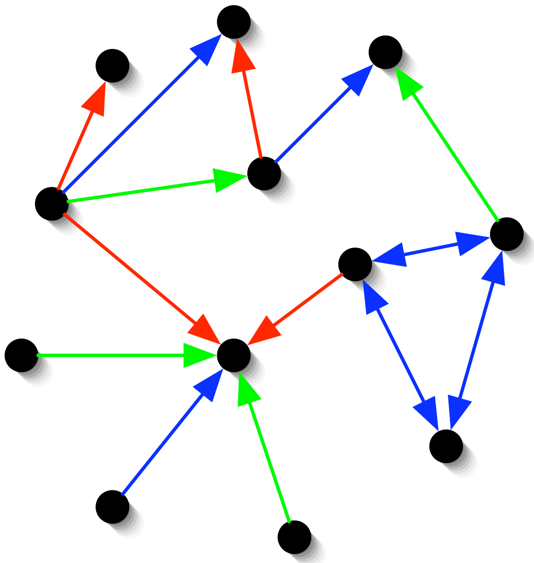
Examples:

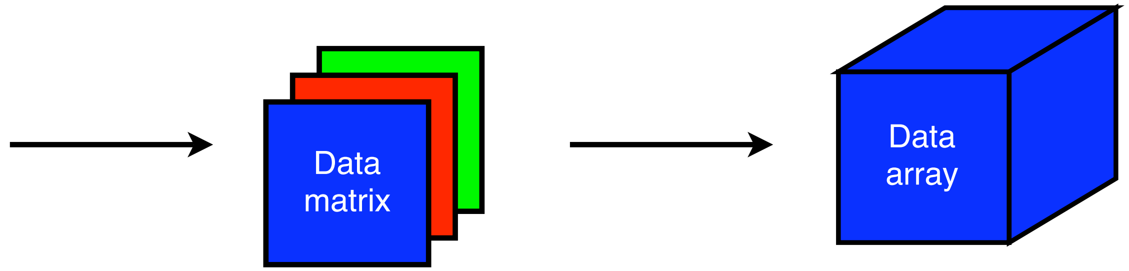
- Latent Semantic Analysis
- Text Analysis (LSI)
- Web search (HITS)
- Clustering

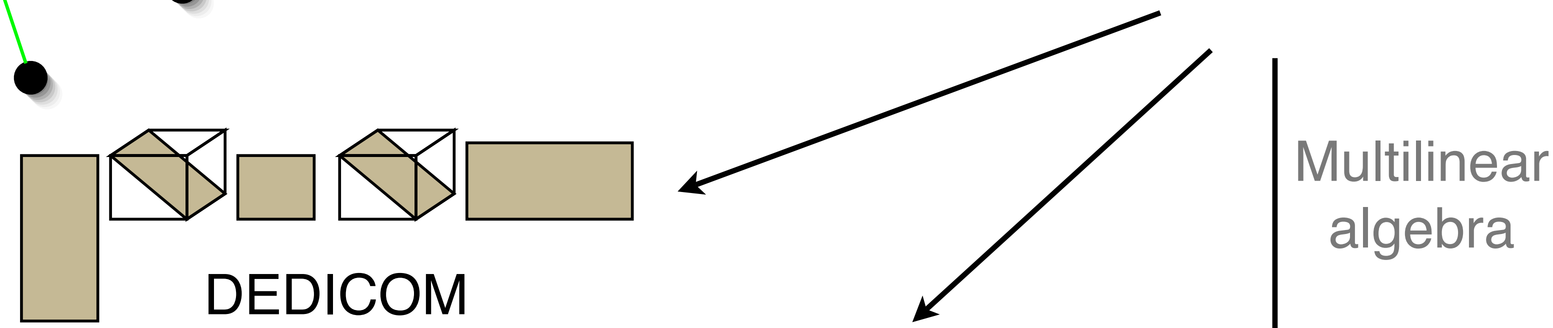
But there may be more useful information in the data!

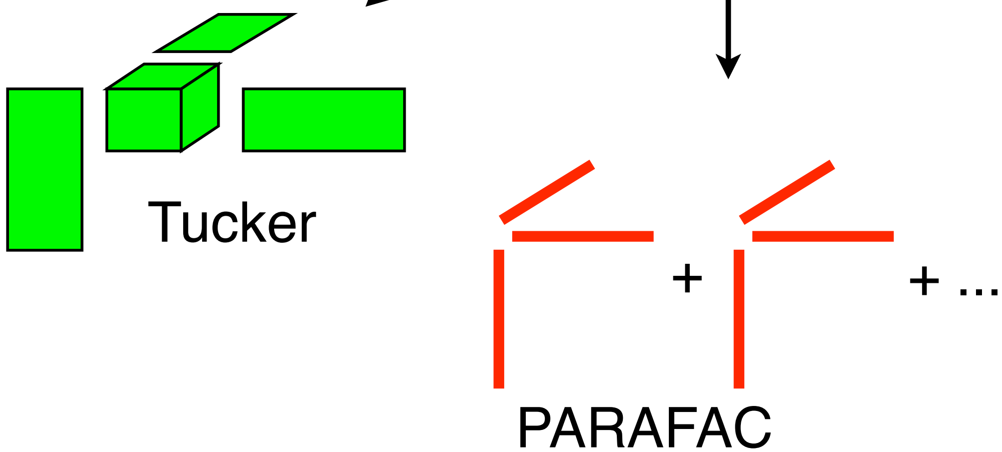
New Paradigm: “Multidimensional Data Mining”

Build a “data array” such that there is a data matrix for each link type.







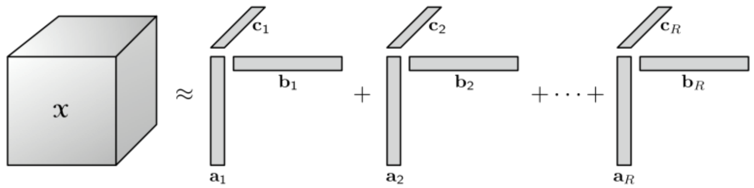


Third dimension offers more explanatory power: uncovers new latent information and reveals subtle relationships

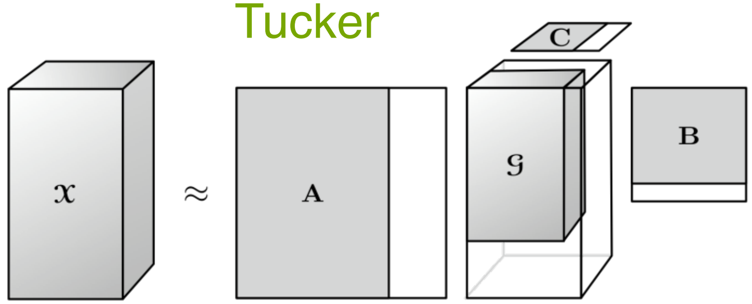
★ Unique data mining capability developed at Sandia

Many Types of Tensor Decompositions

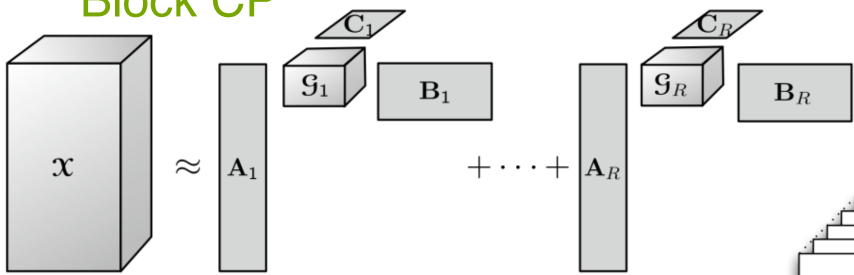
CANDECOMP/PARAFAC (CP)



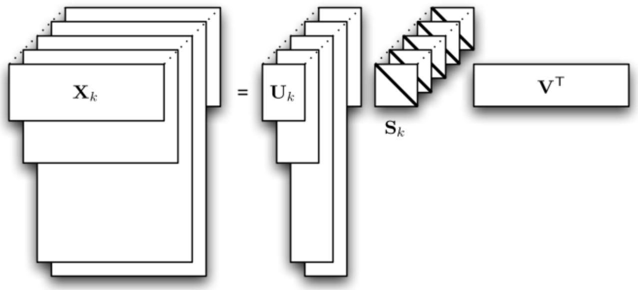
Tucker



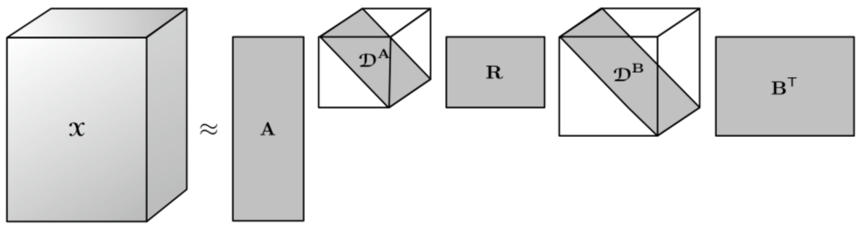
Block CP



PARAFAC2



PARATUCK2 (DEDICOM3)

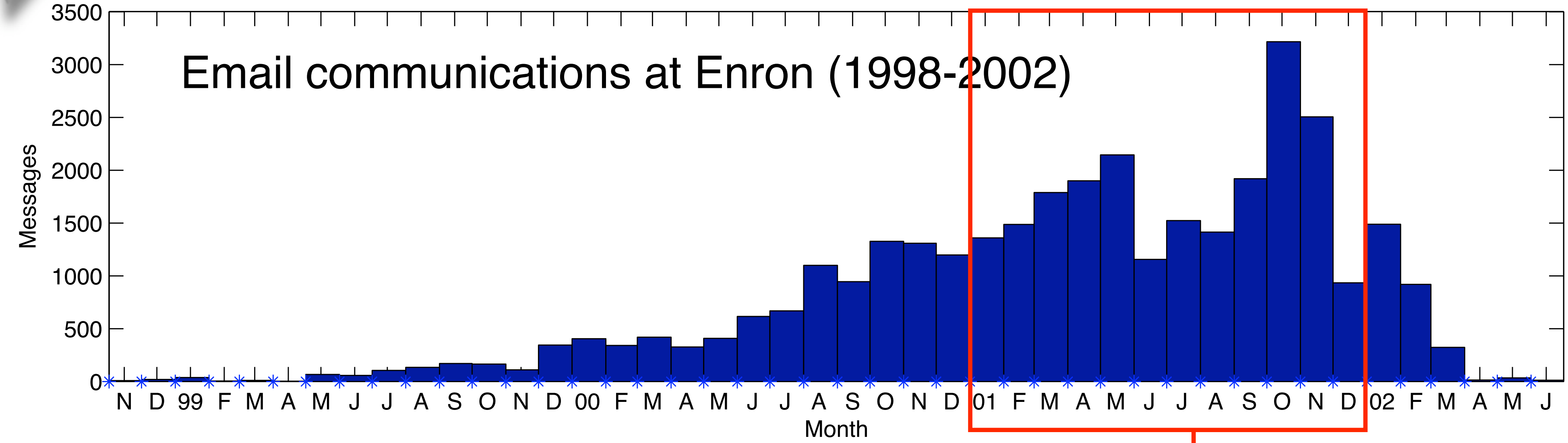


Kolda & Bader, Tensor Decompositions and Applications, SIAM Review, 2009

Each decomposition provides a different interpretation of the data

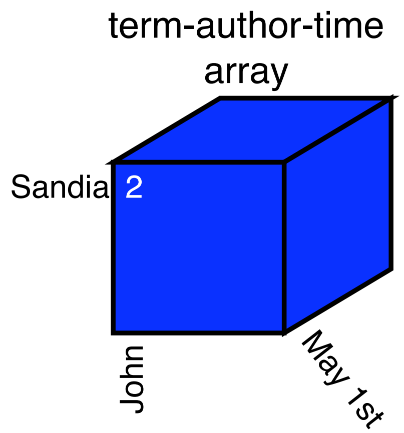
Sandia National Laboratories

Case Study: Discussion Tracking in Email



53,733 messages from 184 employees

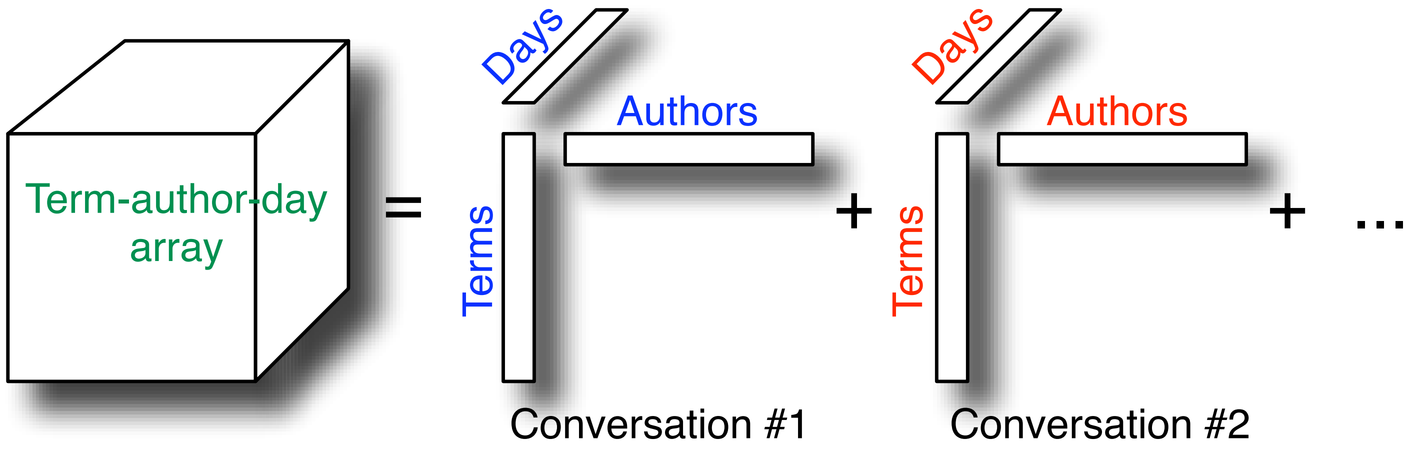
- Situational awareness
- What can we learn from these email conversations?
  - What are the major topics of conversations?
  - Who are the major participants?
  - When are they taking place?

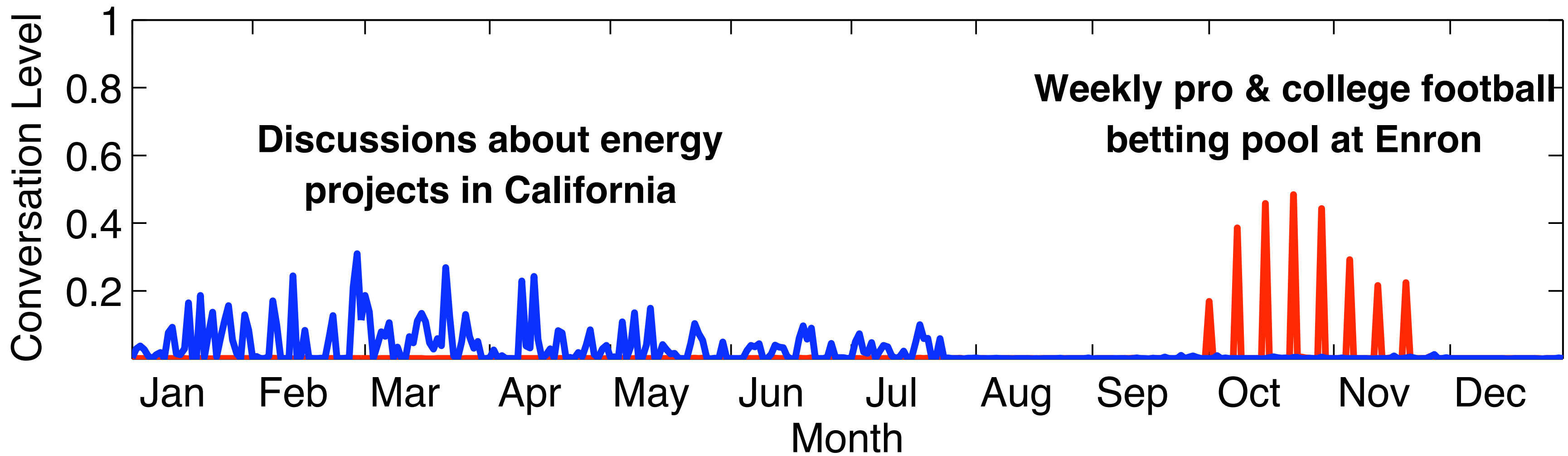


Bader, Berry, Browne, Discussion tracking in Enron email using PARAFAC, in Survey of Text Mining II, 2008.

Sandia National Laboratories

Tensor analysis finds unusual activity by associating terms with people over time



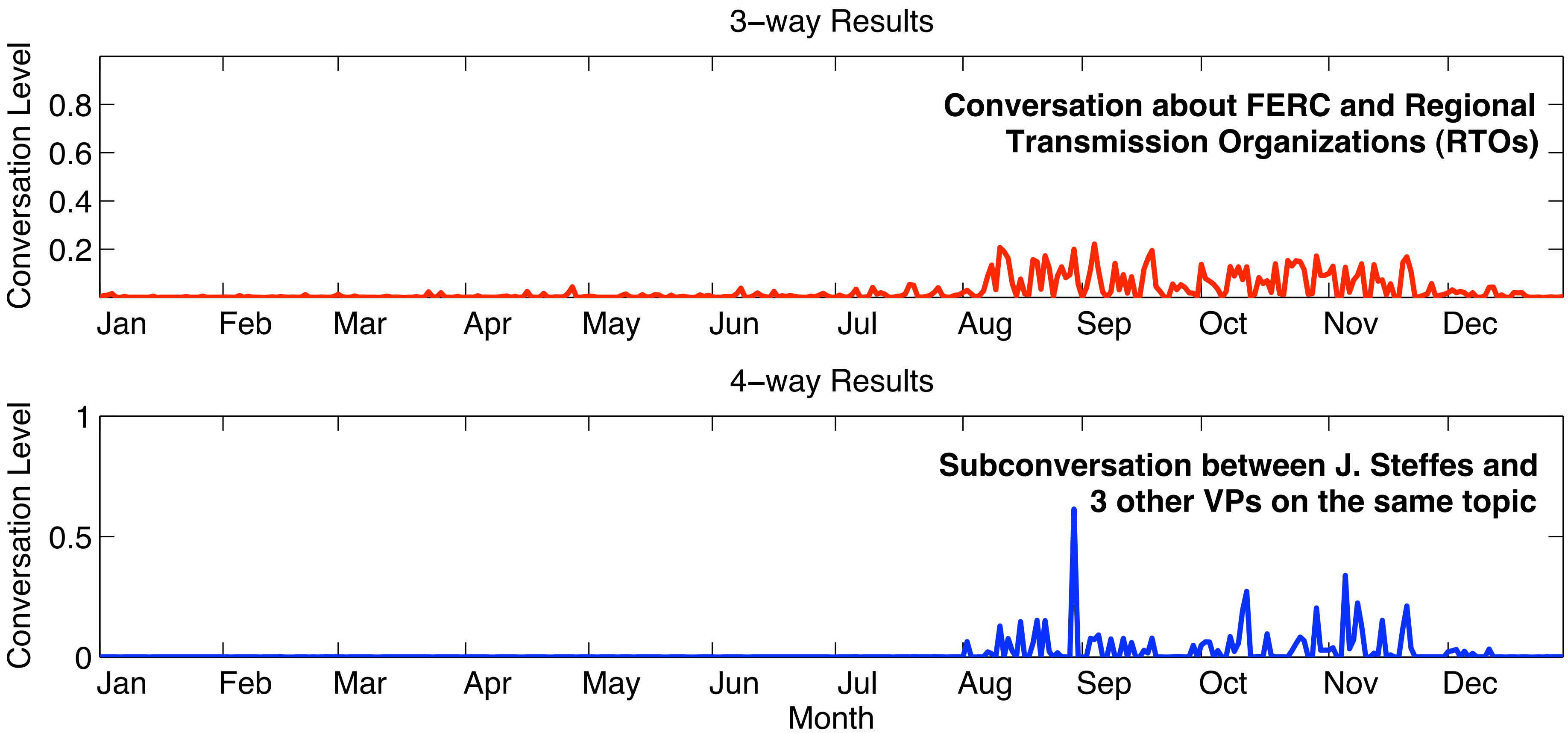


Key terms:	California, power, utilities, energy, utility, governor, market	games, week, missed, picked, prize, wins, scored, upsets
Key authors:	J. Steffes, S. Kean, J. Dasovich, R. Shapiro, P. Allen, ...	A. Pace, L. Campbell, C. Dean

Four-way analysis shows deeper relationships

4-way array: Author x Recipient x Terms x Time

- 4-way analysis may track subconversation already found by 3-way analysis
- Provides context and temporal patterns of social network



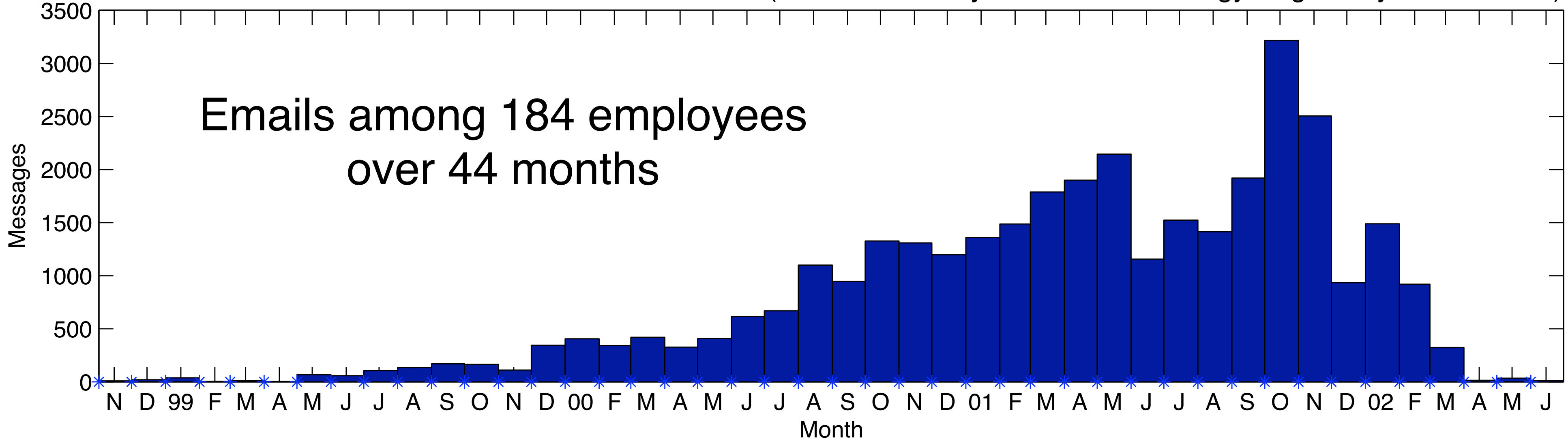
Bader, Berry, Langville, Text analysis using nonnegative matrix/tensor factorizations, in Text Mining, 2009.

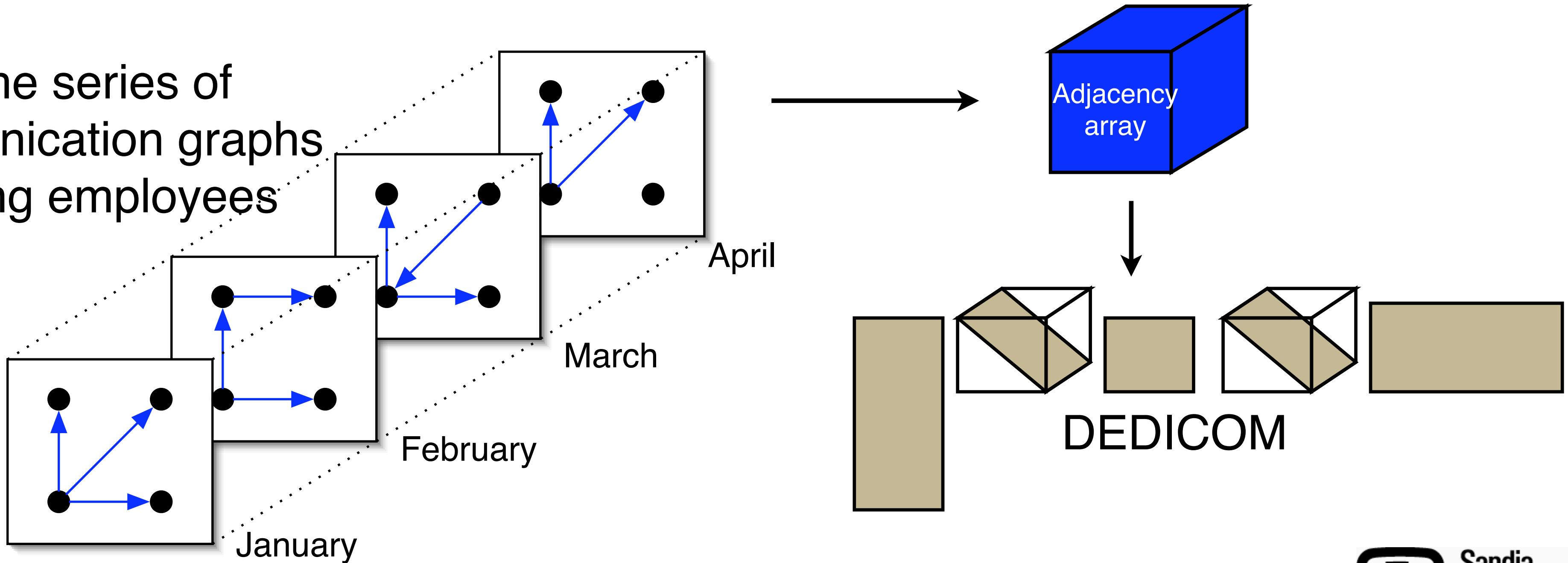
Sandia National Laboratories

Case Study: Pattern Analysis in Email Networks

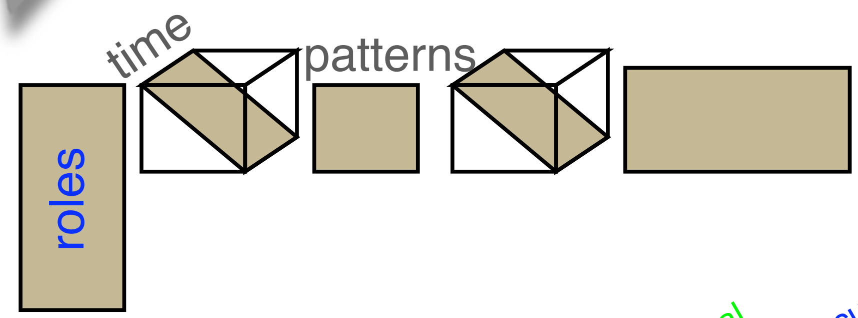
Email communications at Enron (1998-2002)

(data released by U.S. Federal Energy Regulatory Commission)



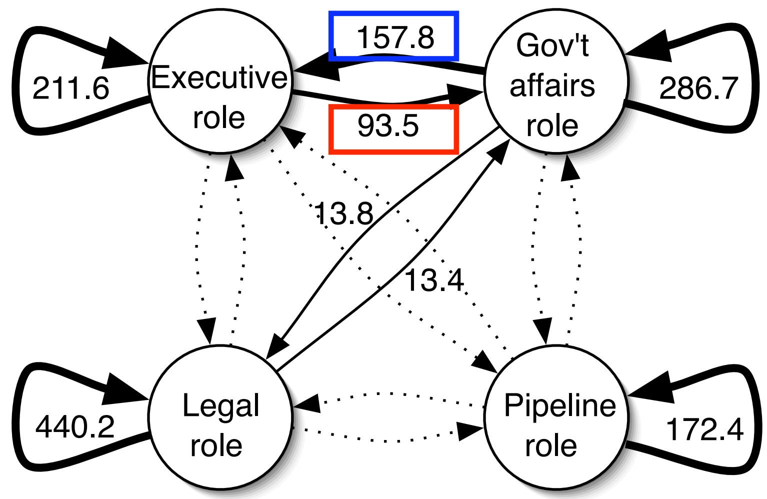


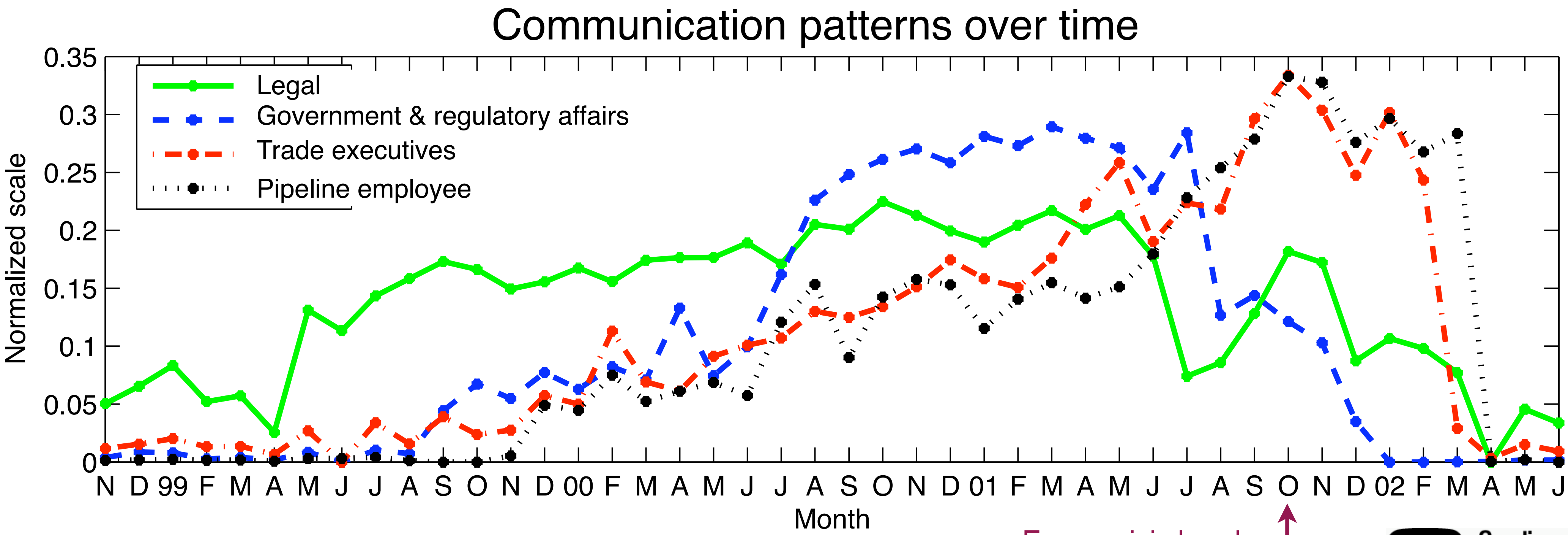
Analysis shows employee roles and communication patterns



	Legal	Executive (gov't affairs)	Executive (trade)	Pipeline
L. Kitchen - President, Enron Online	**0.11**	-0.09	**0.53**	0.00

Soft clustering



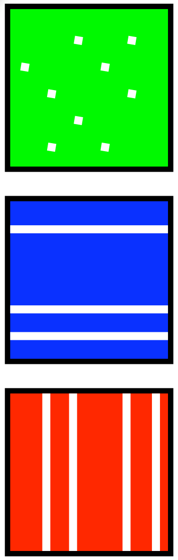


Enron crisis breaks; investigation begins

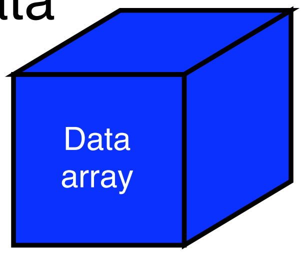
Bader, Harshman, Kolda, Temporal analysis of semantic graphs using ASALSAN, in ICDM 2008.

Sandia National Laboratories

Our algorithms can handle missing data

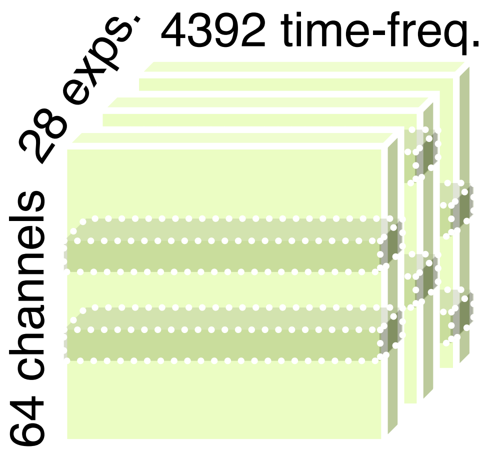


Random or systematic patterns of missing data

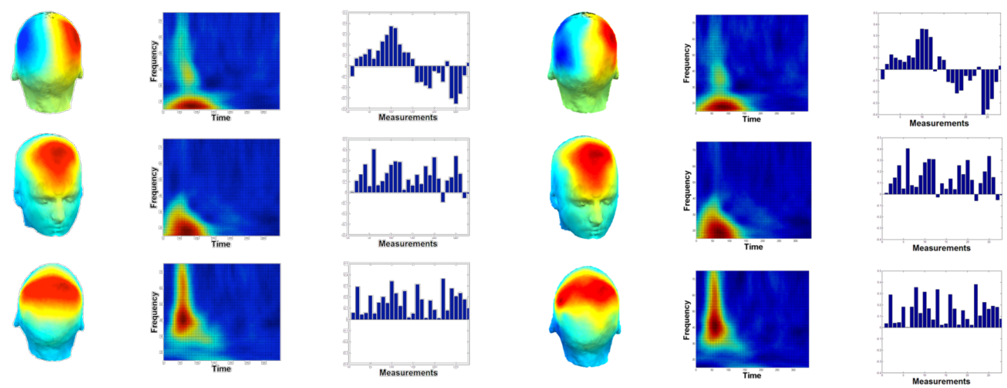


Fit model using derivative-based algorithms

- Simultaneous analysis in 3 ways will fill in the gaps
- Our approach is faster than alternatives
- Specialized algorithm for large-scale problems
  - 500 x 500 x 500 with 99% missing data (1.25M nonzeros)
  - 1000 x 1000 x 1000 with 99.5% missing data (5M nonzeros)



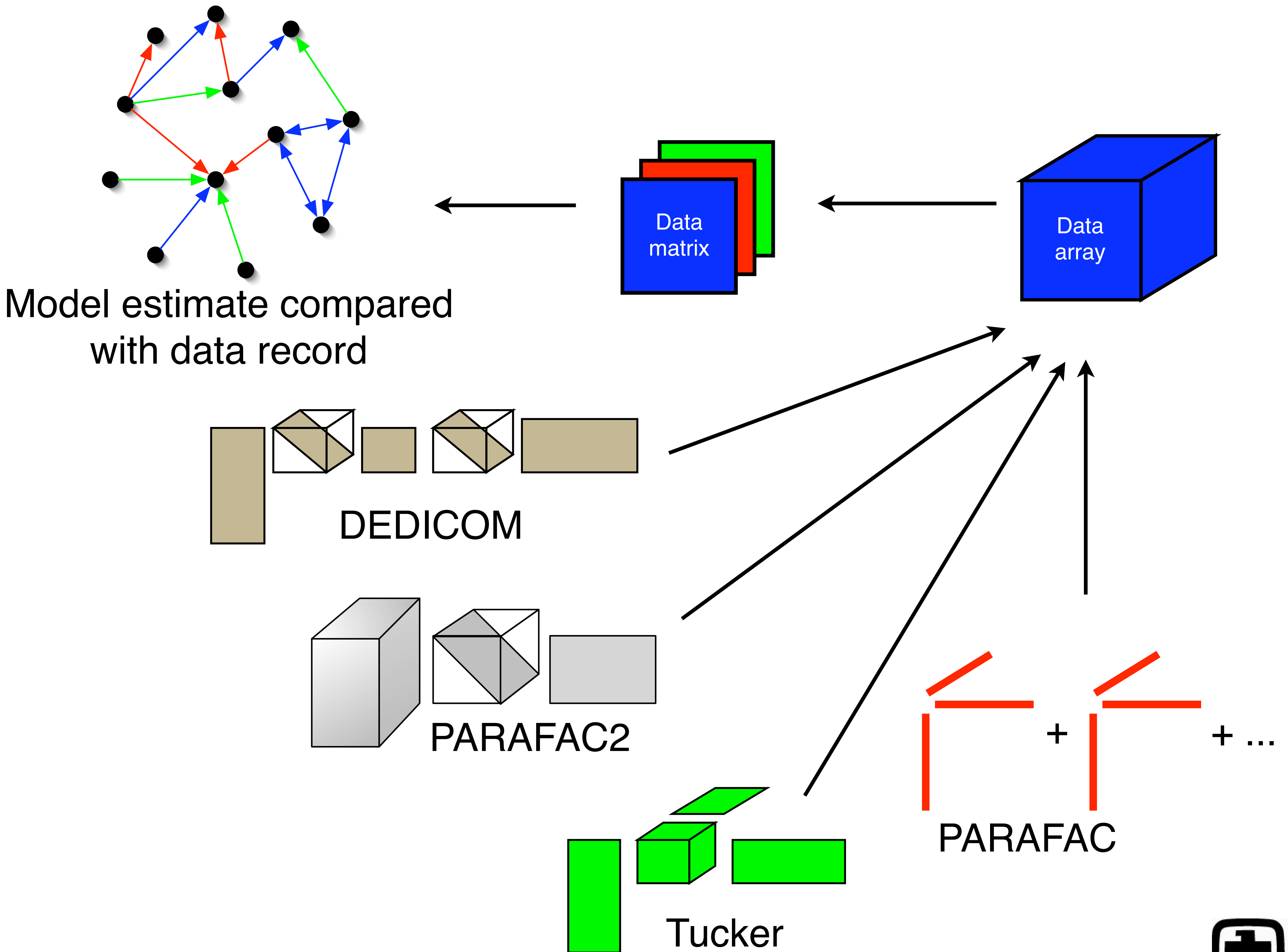
No Missing Data   46% Missing Data



- Acar, Dunlavy, Kolda, Mørup, Scalable Tensor Factorization with Missing Data, SDM2010.
- Acar, Dunlavy, Kolda, Mørup, Scalable Tensor Factorization with Incomplete Data, in revision, 2010.

Sandia National Laboratories

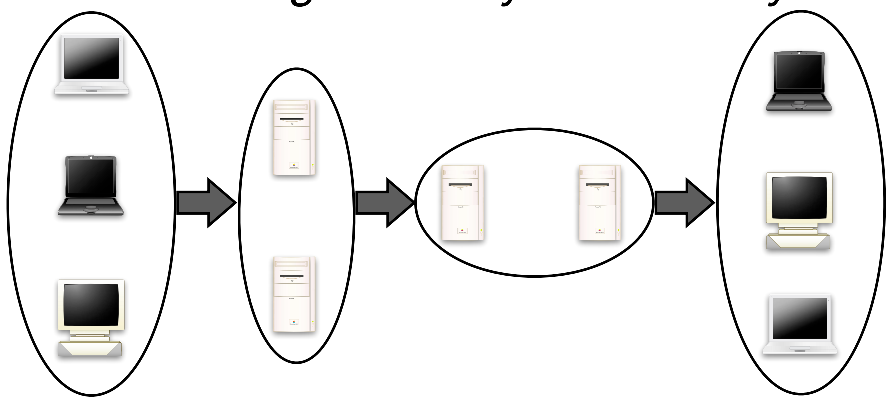
Missing data facilitates another approach to anomaly detection



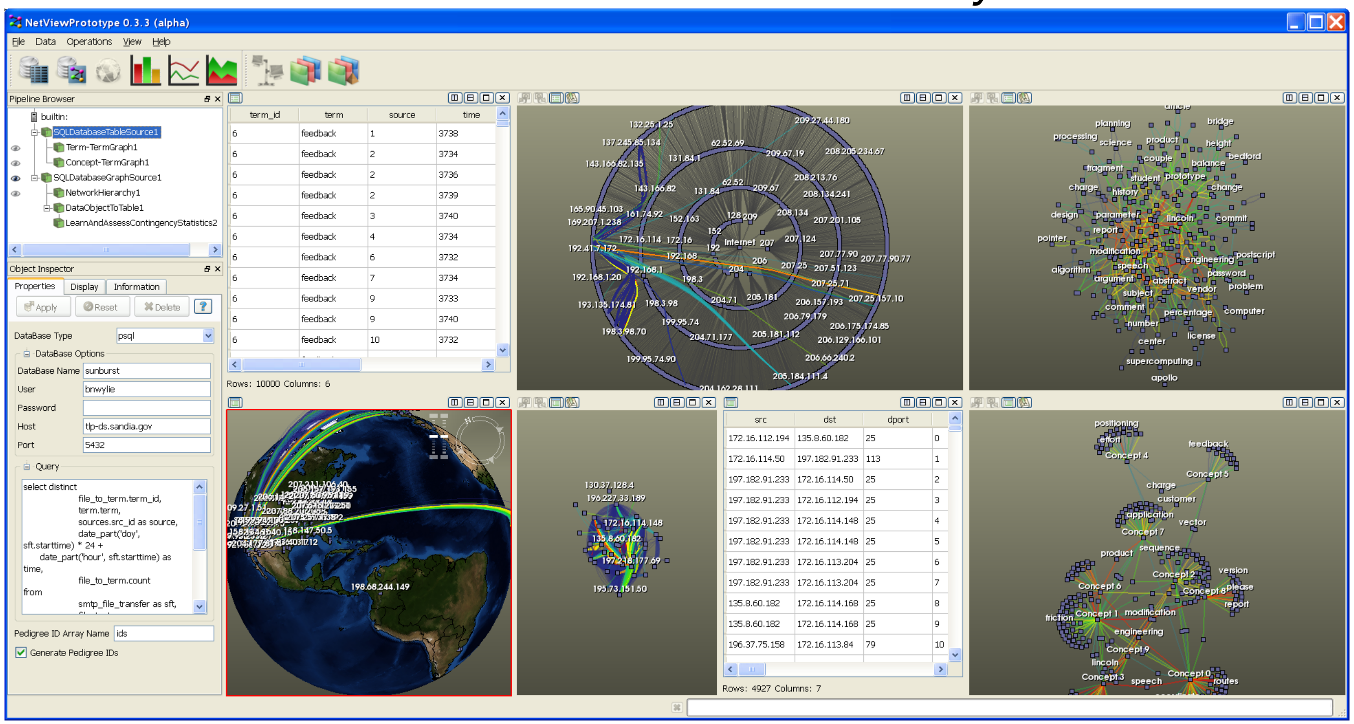
Related Analysis Projects

- Multilingual document analysis and classification
- Uncovering plots buried in text (scenario discovery)
- IP address characterization (trace route analysis)
- Network traffic analysis (cyber, phone)
- Cyber data exfiltration analysis
- Link prediction
- Higher-order web link analysis

Clustering nodes by their activity

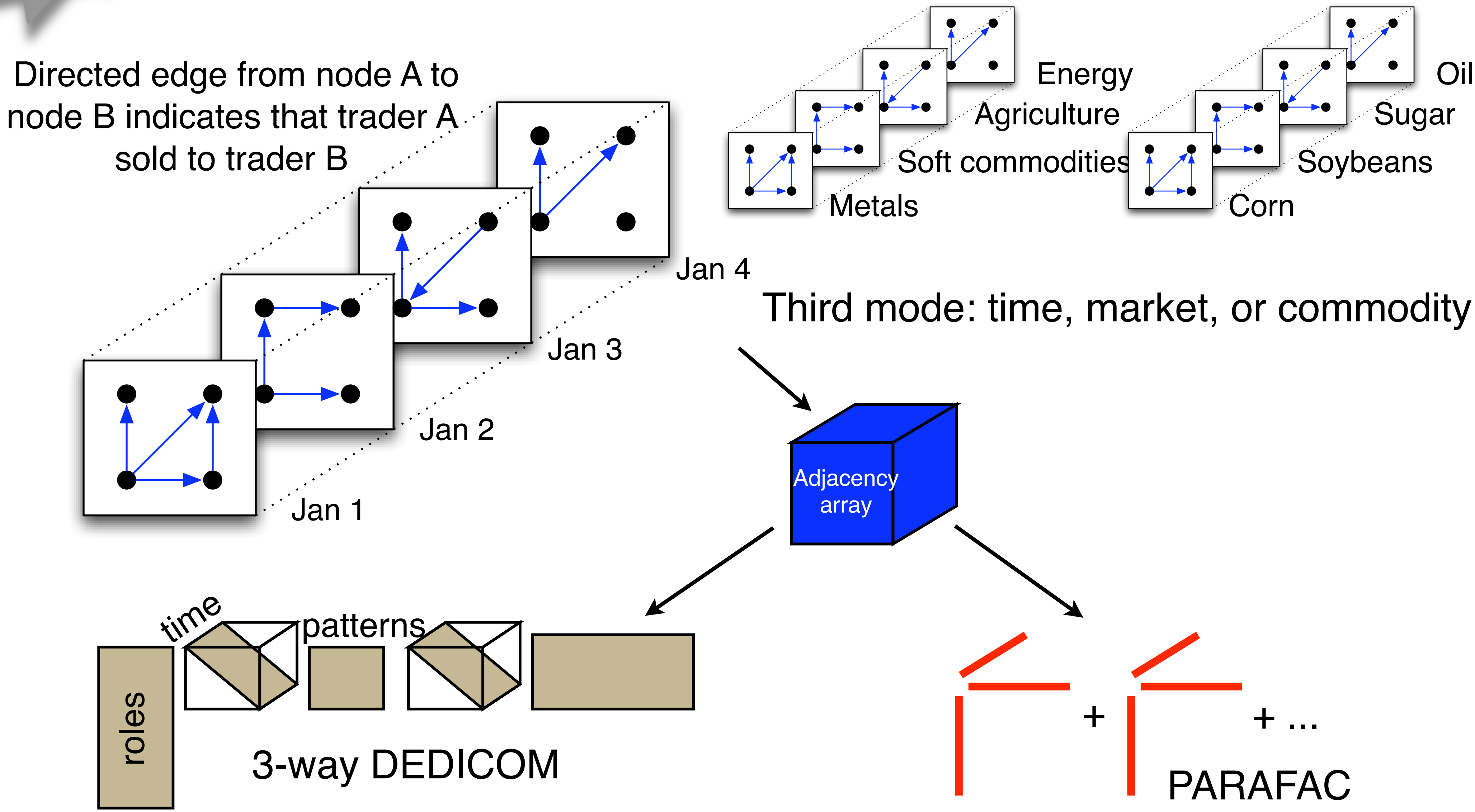


Network traffic and content analysis



Sandia National Laboratories

Preliminary Ideas for Trading Networks



- Soft clustering of traders by their activity
- Aggregate trading patterns among clusters
- Behavior over time or by market

- Traders characterized by their "authority"
- Patterns in time or by market