SAND2006-3425C

Improvements to 3-way DEDICOM for Applications in Social Network Analysis

Brett Bader*, Richard Harshman** & Tamara Kolda*
*Sandia National Laboratories
**University of Western Ontario

TRICAP
June 5, 2006

National Nuclear Security Administration

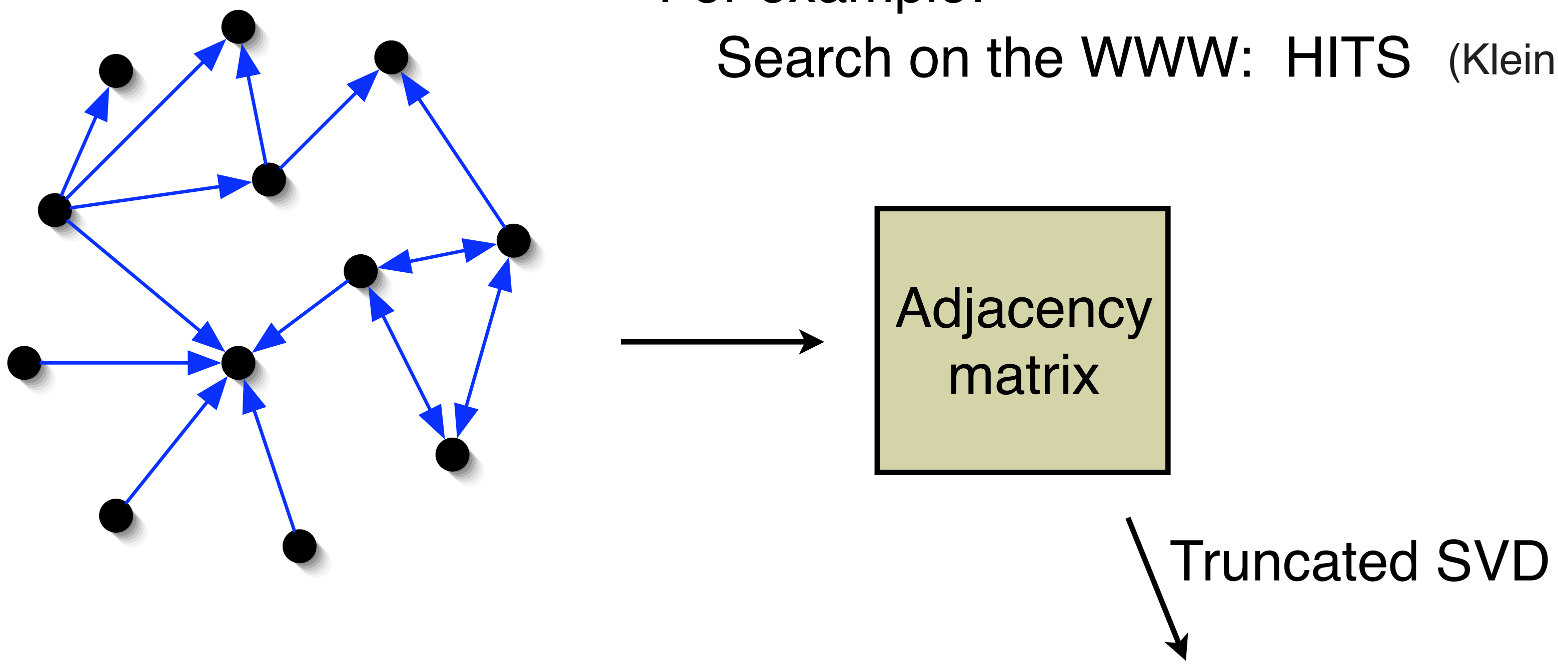
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Common Graph Analysis Technique

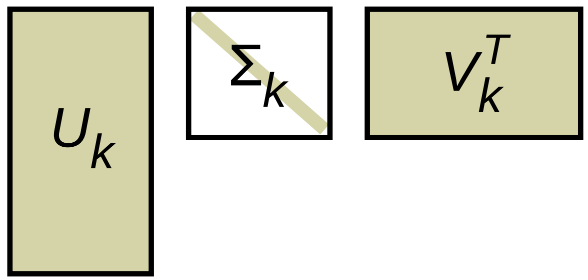
For example:

Search on the WWW: HITS (Kleinberg, 1998)



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

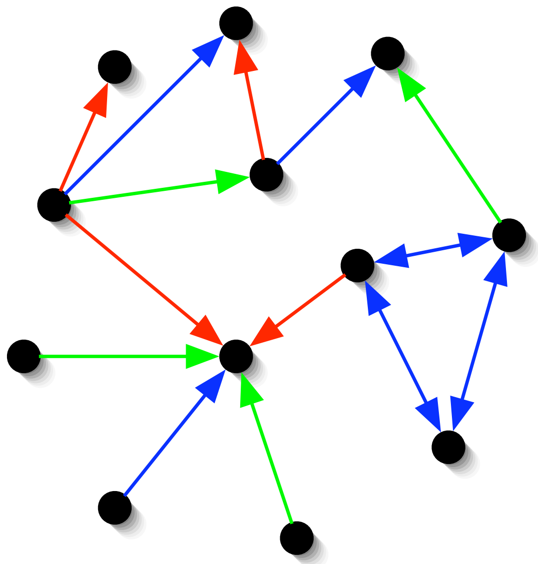
$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$



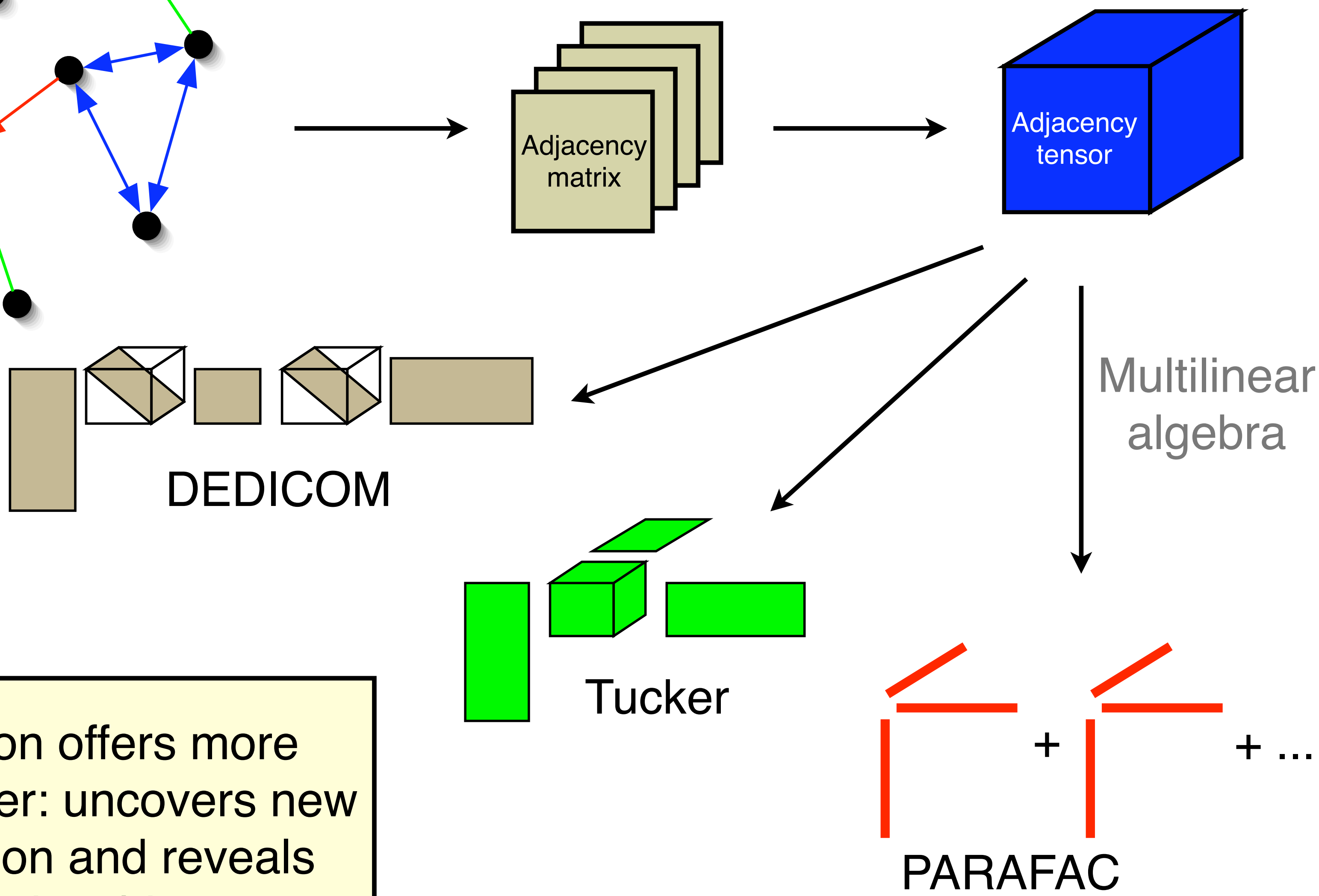
But we may have lost critical information by ignoring edge metadata!

Sandia National Laboratories

New Paradigm: "Multidimensional Data Mining"



Build an "adjacency tensor" such that there is an adjacency matrix for each type of edge.

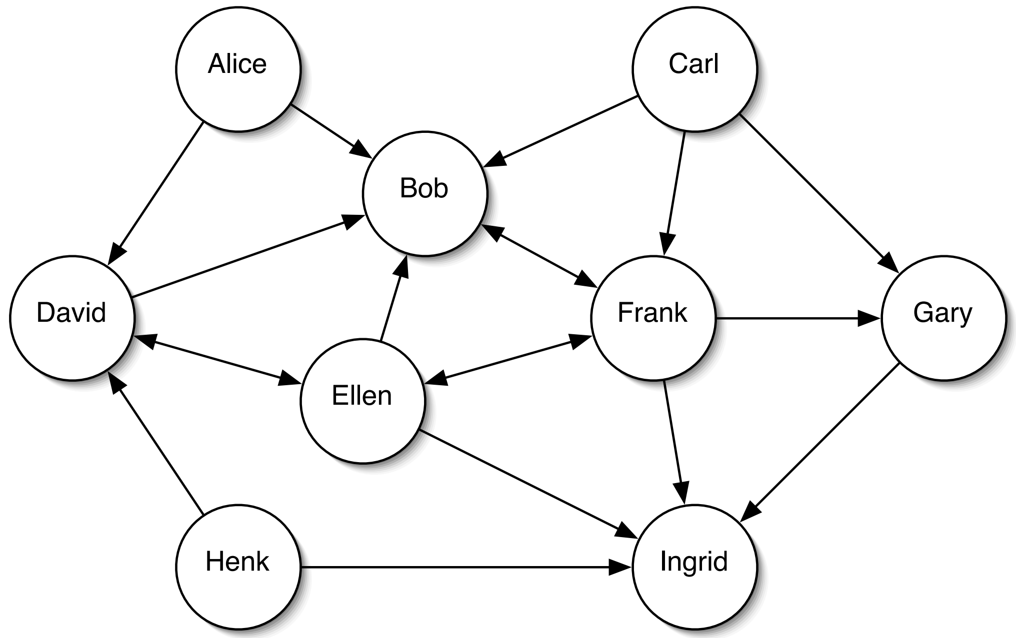


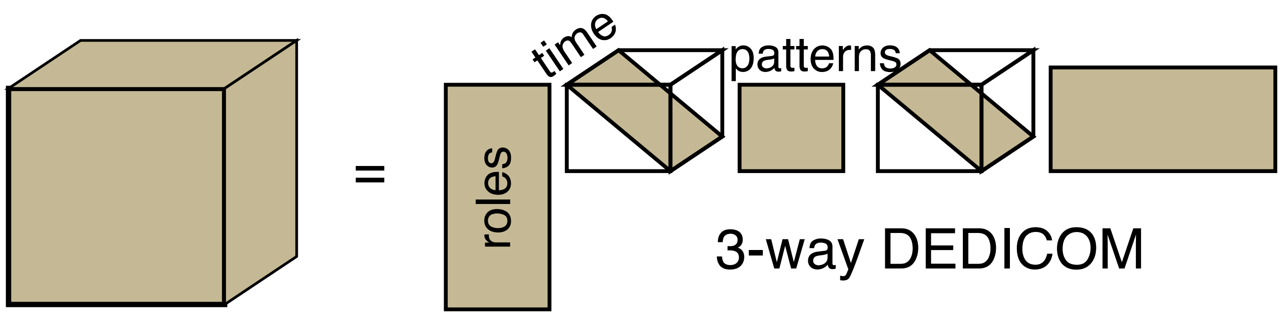
Third dimension offers more explanatory power: uncovers new latent information and reveals subtle relationships

Sandia National Laboratories

Using Tensors for Graph Analysis

Use 3-way DEDICOM to analyze complex social networks that change over time



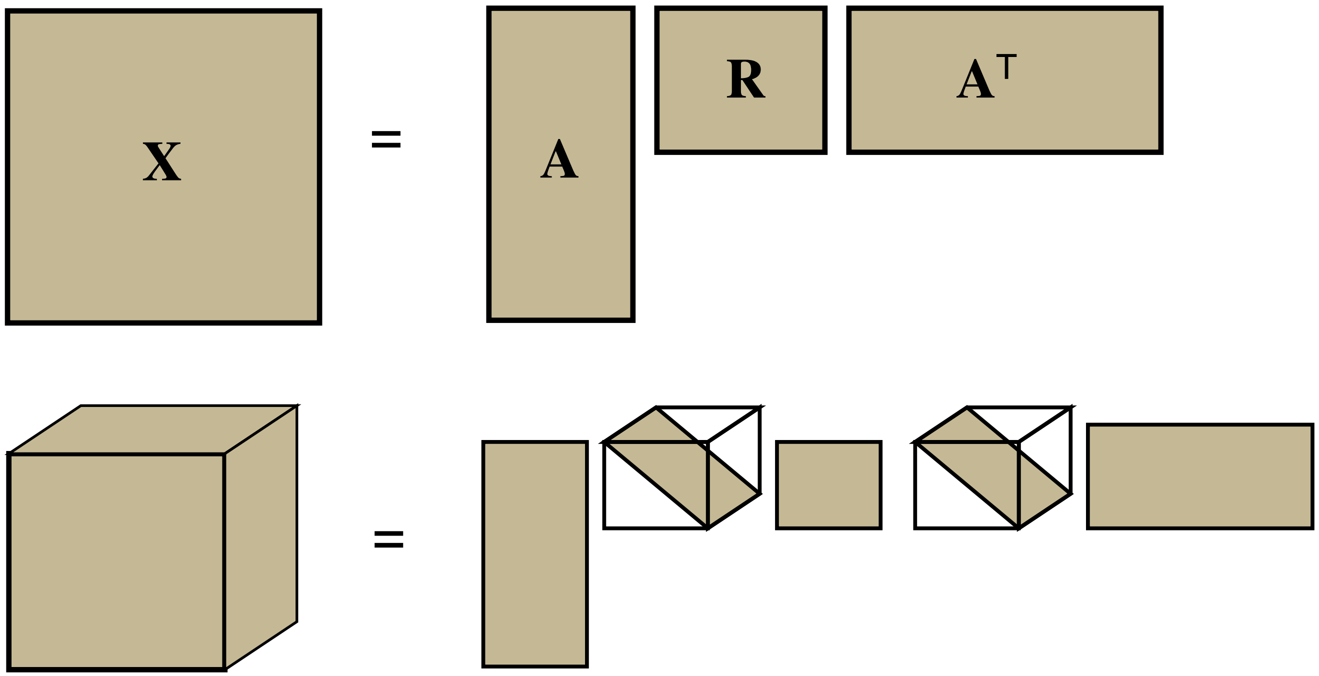


Sandia National Laboratories

DEDICOM

- DEcomposition into DIrectional COMponents
- Introduced in 1978 by Harshman
- Part of the family of models called PARATUCK2
- Past applications
  - Study asymmetries in telephone calls among cities
  - Marketing research
    - car “switching” - car owners and what they buy next
    - free associations of words (e.g., shampoo: “body” evokes “fullness”)
  - Asymmetric measures of world trade (import/export)
- Variations
  - Three-way DEDICOM
  - Constrained DEDICOM
  - Skew-symmetric data

DEDICOM Algorithms



Solve by Alternating Least Squares

- Generalized Takane method
- Kiers' method
- New algorithm

Mathematical Notation

- Scalars a
- Vectors $\mathbf{a}$
- Matrices $\mathbf{A}$
- Tensors (3-way array) $\mathcal{D}$ $\mathcal{X}$
  - frontal slices of $\mathcal{X}$: $\mathbf{X}_i$

- Special symbols
  - Kronecker product

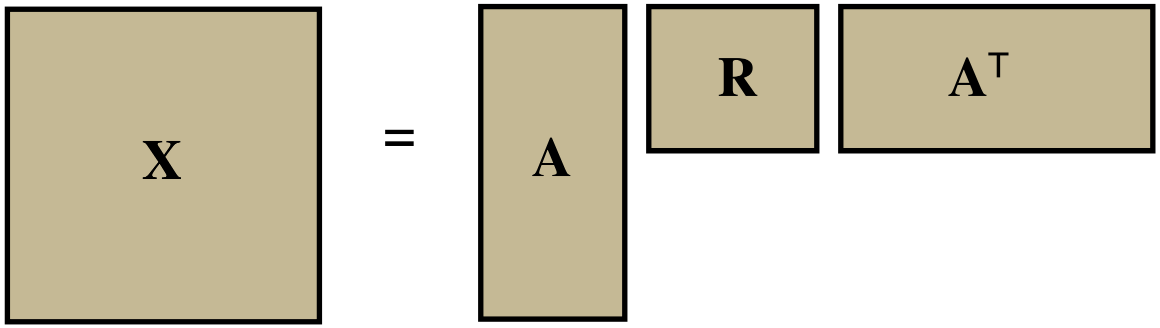
$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

  - Hadamard product (elementwise)

$$\mathbf{A} * \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & \dots & a_{1n}b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & \dots & a_{mn}b_{mn} \end{bmatrix}$$

DEDICOM

$$\mathbf{X} = \mathbf{A}\mathbf{R}\mathbf{A}^T + \mathbf{E}$$
$$\mathbf{X} \approx \mathbf{A}\mathbf{R}\mathbf{A}^T$$



$$\min_{\mathbf{A},\mathbf{R}} \left\| \mathbf{X} - \mathbf{A}\mathbf{R}\mathbf{A}^T \right\|_F^2$$
s.t. $\mathbf{A}$ orthogonal

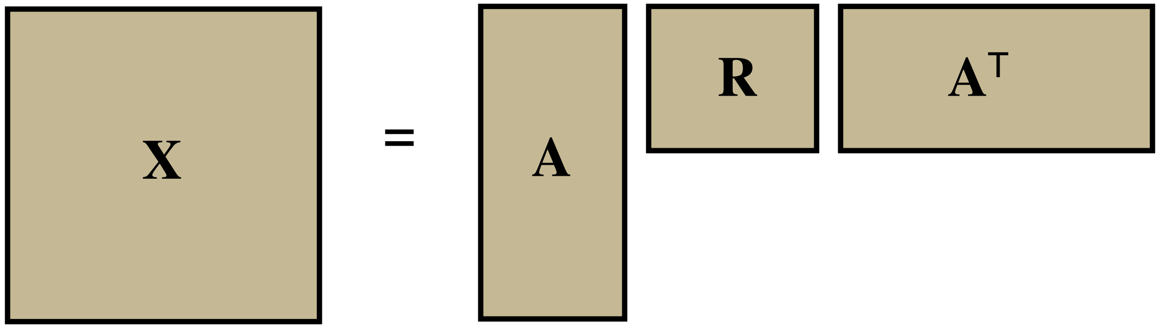
- $\mathbf{A}$ (n x p) is a matrix of loadings or weights
- $\mathbf{R}$ (p x p) is a matrix that captures asymmetric relationships
- $\mathbf{A}$ is not unique
  - $\mathbf{A}$ can be transformed with no loss of fit to the data
  - Nonsingular transformation $\mathbf{Q}$:
  -
  $$\mathbf{A}\mathbf{R}\mathbf{A}^T = (\mathbf{A}\mathbf{Q})(\mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-T})(\mathbf{A}\mathbf{Q})^T$$
  -
  - Usually “fix” A with some standard rotation (e.g., VARIMAX)

Generalized Takane Method

(Takane, 1985; Kiers et al., 1990)

$$\mathbf{X} \approx \mathbf{A}\mathbf{R}\mathbf{A}^T$$

$$\min_{\mathbf{A},\mathbf{R}} \left\| \mathbf{X} - \mathbf{A}\mathbf{R}\mathbf{A}^T \right\|_F^2$$



Loss function: $\sigma(\mathbf{A},\mathbf{R}) = \left\| \mathbf{X} - \mathbf{A}\mathbf{R}\mathbf{A}^T \right\|_F^2$

Minimizing σ wrt $\mathbf{A}$, for fixed $\mathbf{R}$,
is equivalent to maximizing: $f(\mathbf{A}) = tr(\mathbf{A}^T\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{A})$

Compute $\mathbf{A}$ via Gram-Schmidt
orthonormalization: $(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{A} + \mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X}\mathbf{A} + 2\alpha\mathbf{A})$
where α is > largest eigenvalue of
symmetric part of $(-\mathbf{X} \otimes \mathbf{A}^T\mathbf{X}\mathbf{A})$

Practical method:

- Compute an update for $\mathbf{A}$ using $\alpha = 0$
- check if $f(\mathbf{Anew}) > f(\mathbf{A})$
- If not, compute $\mathbf{A}$ using nonzero α

New Algorithm

Solving for $\mathbf{A}$:

Stack data "side by side"

$$\begin{aligned}(\mathbf{X} \quad \mathbf{X}^T) &= (\mathbf{ARA}^T \quad \mathbf{AR}^T\mathbf{A}^T) \\ &= \mathbf{A}\left((\mathbf{R} \quad \mathbf{R}^T)\begin{pmatrix}\mathbf{A}^T & 0 \\ 0 & \mathbf{A}^T\end{pmatrix}\right)\end{aligned}$$

...and solve extended LS problem, i.e., $\min_{\mathbf{A}} \| \mathbf{Y} - \mathbf{AB} \|_F^2$

$$\mathbf{A}_{new} \leftarrow (\mathbf{X} \quad \mathbf{X}^T)\left((\mathbf{R} \quad \mathbf{R}^T)\begin{pmatrix}\mathbf{A}^T & 0 \\ 0 & \mathbf{A}^T\end{pmatrix}\right)^{\dagger}$$

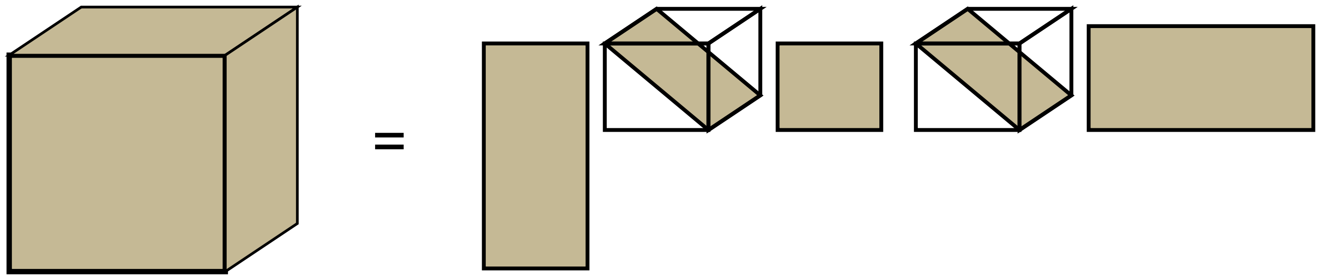
or

$$\mathbf{A}_{new} = \left(\mathbf{XAR}^T + \mathbf{X}^T\mathbf{AR}\right)\left(\mathbf{R}(\mathbf{A}^T\mathbf{A})\mathbf{R}^T + \mathbf{R}^T(\mathbf{A}^T\mathbf{A})\mathbf{R}\right)^{-1}.$$

Solving for $\mathbf{R}$:

$$\mathbf{R}_{new} = \mathbf{A}^{\dagger}\mathbf{X}(\mathbf{A}^T)^{\dagger}.$$ (same as before)

Three-way DEDICOM



$$\mathbf{X}_i = \mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\mathbf{A}^T + \mathbf{E}_i \quad \text{for } i = 1, \ldots, m,$$

$$\min_{\mathbf{A},\mathbf{R},\mathcal{D}} \sum_{i=1}^{m} \| \mathbf{X}_i - \mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\mathbf{A}^T \|_F^2$$

- $\mathbf{A}$ (n x p) is matrix of loadings or weights (not necessarily orthogonal)
- $\mathbf{R}$ (p x p) is a matrix that captures asymmetric relationships
- D (p x p x m) gives the weights of the columns of A for each level in third mode
- Unique solution with enough levels of $\mathbf{X}$

Kiers' Algorithm

(Kiers, 1993)

$$\min_{\mathbf{A},\mathbf{R},\boldsymbol{\mathcal{D}}} \sum_{i=1}^{m} \| \mathbf{X}_i - \mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\mathbf{A}^T \|_F^2$$

<u>Alternating Least Squares</u>

1) Column-wise minimization to find $\mathbf{A}$

potentially slow convergence

2) Least-squares problem for $\mathbf{R}$

$$\text{minimize: } f(\mathbf{R}) = \left\| \begin{pmatrix} \text{Vec}(\mathbf{X}_1) \\ \vdots \\ \text{Vec}(\mathbf{X}_m) \end{pmatrix} - \begin{pmatrix} \mathbf{A}\mathbf{D}_1 \otimes \mathbf{A}\mathbf{D}_1 \\ \vdots \\ \mathbf{A}\mathbf{D}_m \otimes \mathbf{A}\;\mathbf{D}_m \end{pmatrix} \text{Vec}(\mathbf{R}) \right\|$$

$$\text{Vec}(\mathbf{R}) = \left(\sum_{i=1}^{m} (\mathbf{D}_i\mathbf{A}^T\mathbf{A}\mathbf{D}_i) \otimes (\mathbf{D}_i\mathbf{A}^T\mathbf{A}\mathbf{D}_i) \right)^{-1} \sum_{i=1}^{m} \text{Vec}(\mathbf{D}_i\mathbf{A}^T\mathbf{X}_i\mathbf{A}\mathbf{D}_i)$$

3) Element-wise minimization to find $\mathbf{D}$

potentially slow convergence

Sandia National Laboratories

New Algorithm - updating A

Solving for $\mathbf{A}$:

$$\begin{pmatrix}\mathbf{X}_1 & \mathbf{X}_1^T & \cdots & \mathbf{X}_m & \mathbf{X}_m^T\end{pmatrix} = \mathbf{A}\begin{pmatrix}\mathbf{D}_1\mathbf{R}\mathbf{D}_1 & \mathbf{D}_1\mathbf{R}^T\mathbf{D}_1 & \cdots & \mathbf{D}_m\mathbf{R}\mathbf{D}_m & \mathbf{D}_m\mathbf{R}^T\mathbf{D}_m\end{pmatrix}\left(\mathbf{I}_{2m} \otimes \mathbf{A}^T\right)$$

$$\mathbf{A} = \left[\sum_{i=1}^{m}\left(\mathbf{X}_i\mathbf{A}\mathbf{D}_i\mathbf{R}^T\mathbf{D}_i + \mathbf{X}_i^T\mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\right)\right]\left[\sum_{i=1}^{m}(\mathbf{B}_i + \mathbf{C}_i)\right]^{-1}$$

$$\begin{aligned}\mathbf{B}_i &\equiv \mathbf{D}_i\mathbf{R}\mathbf{D}_i(\mathbf{A}^T\mathbf{A})\mathbf{D}_i\mathbf{R}^T\mathbf{D}_i,\\ \mathbf{C}_i &\equiv \mathbf{D}_i\mathbf{R}^T\mathbf{D}_i(\mathbf{A}^T\mathbf{A})\mathbf{D}_i\mathbf{R}\mathbf{D}_i.\end{aligned}$$

New Algorithm - updating D

$$\min_{\mathbf{D}_i} \| \mathbf{X}_i - \mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\mathbf{A}^T \|_F^2$$

Solving for $\mathbf{D}$:

Use Newton's method to solve the optimization problem

$$d_{new} = d - H^{-1}g$$

$$g_k = -\sum_{i,j} \left[2(\mathbf{X} - \mathbf{ADRDA}^T) * (\mathbf{ADr}_k\mathbf{a}_k^T + \mathbf{a}_k\mathbf{r}_{k,:}\mathbf{DA}^T) \right]_{i,j}$$

$$h_{st} = -2\sum_{i,j} \Big[(\mathbf{X} - \mathbf{ADRDA}^T) * (\mathbf{a}_s r_{st}\mathbf{a}_t^T + \mathbf{a}_t r_{ts}\mathbf{a}_s^T) - (\mathbf{ADr}_s\mathbf{a}_s^T + \mathbf{a}_s\mathbf{r}_{s:}\mathbf{DA}^T) * (\mathbf{ADr}_t\mathbf{a}_t^T + \mathbf{a}_t\mathbf{r}_{t:}\mathbf{DA}^T) \Big]_{i,j}$$

Use compression

QR factorization: $\mathbf{A} = \mathbf{Q}\tilde{\mathbf{A}},$

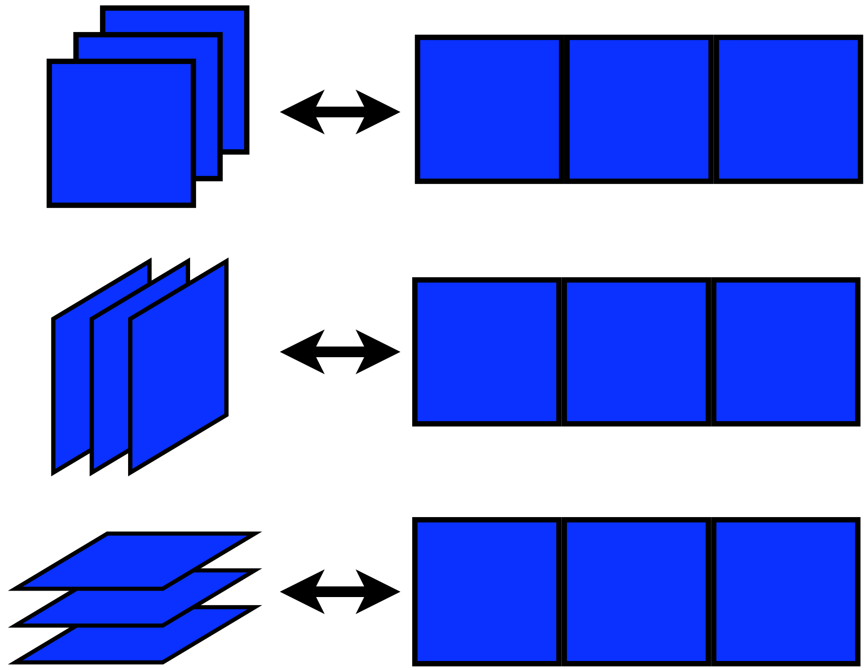
$$\min_{\mathbf{D}_i} \left\| \mathbf{Q}^T\mathbf{X}_i\mathbf{Q} - \tilde{\mathbf{A}}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\tilde{\mathbf{A}}^T \right\|_F^2$$

Smaller problem

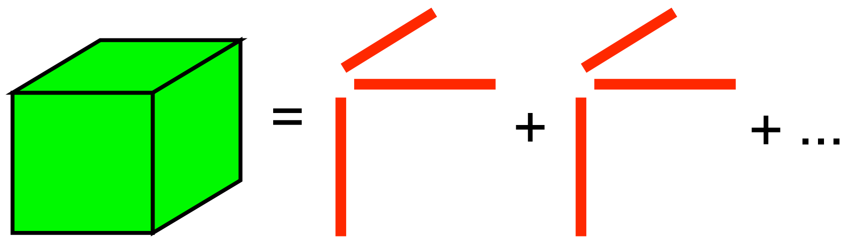
MATLAB Tensor Toolbox

(Bader and Kolda)

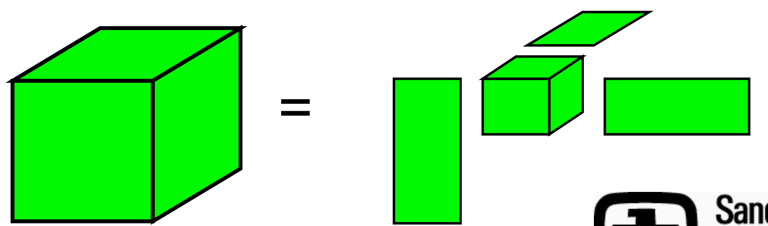
- Toolbox extends functionality of Matlab's MDA datatype:
  - Basic operations
  - Convert to/from a matrix
  - Multiplication
    - Tensor
    - Matrix
    - Vector
- Facilitates rapid prototyping of algorithms
  - PARAFAC/CANDECOMP
  - Tucker
  - DEDICOM
- Extensions for a sparse tensor format (in development)



$$\mathcal{B} = \mathcal{A} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}$$



Note: not intended to replace
Andersson & Bro's N-way Toolbox



http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox

Sandia National Laboratories

New `sparse_tensor` Class

```
sparse_tensor class
```

- Coordinate based storage: (i,j,k) indices & values
  - no maximum order
- Implements many of the same functions as `tensor` class
- Reshape and permute operations handled implicitly with index
- Row / column / slice operations are easy
- Implemented with built-in functions where possible
- Careful to avoid explicit referencing (e.g., `A(i,j,k)`)

Large-scale Sparse Tensors

- 1000 x 1000 x 1000 tensor is huge (1 billion entries!)

- Even worse for higher order tensors

- Dense tensor implementation not possible
  - Storage!
  - `permute` becomes prohibitively expensive
    - At the heart of many operations (e.g., n-mode product)

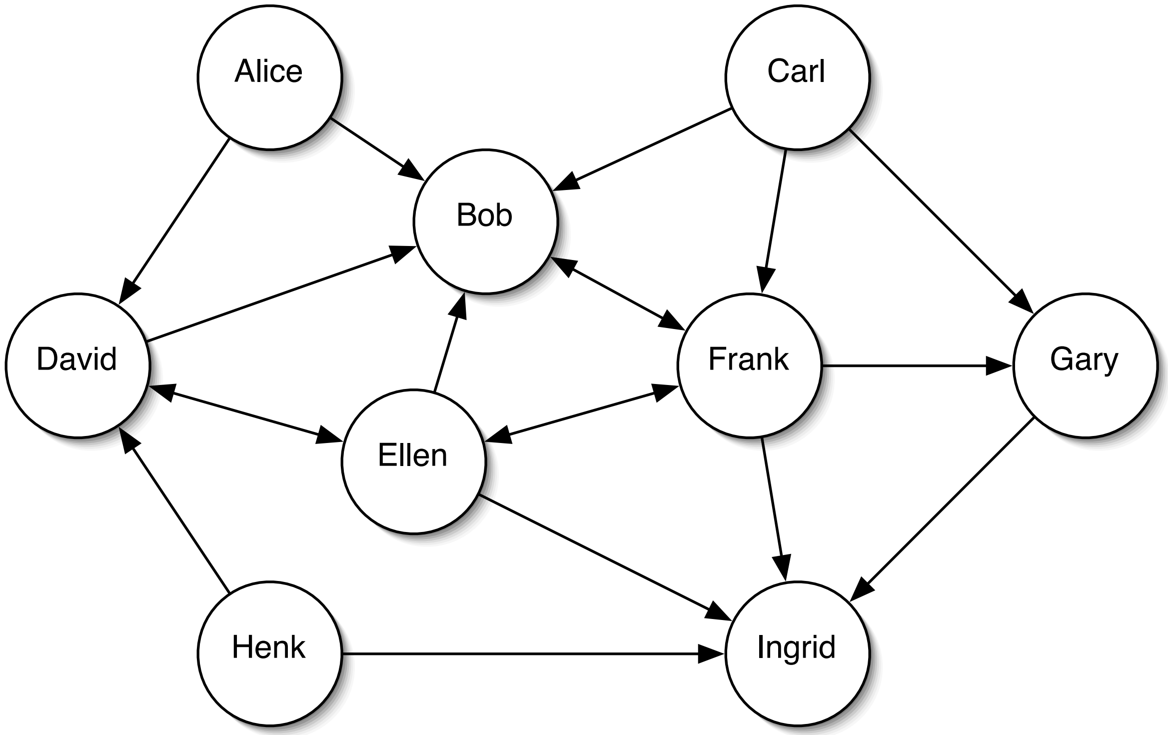
- Need a sparse implementation that is efficient for multiplication over all n modes (i.e., fast permute and reshape)

Sandia National Laboratories

Example Performance

- 73,000 x 73,000 x 40,000 sparse tensor
- 469,000 nonzeros (out of 200 trillion entries, so *very* sparse)
- Compute approximation to rank-10 PARAFAC model
  - 12 minutes on a laptop
  - 329 power method iterations
  - 987 calls to `ttv`

Social Network Analysis

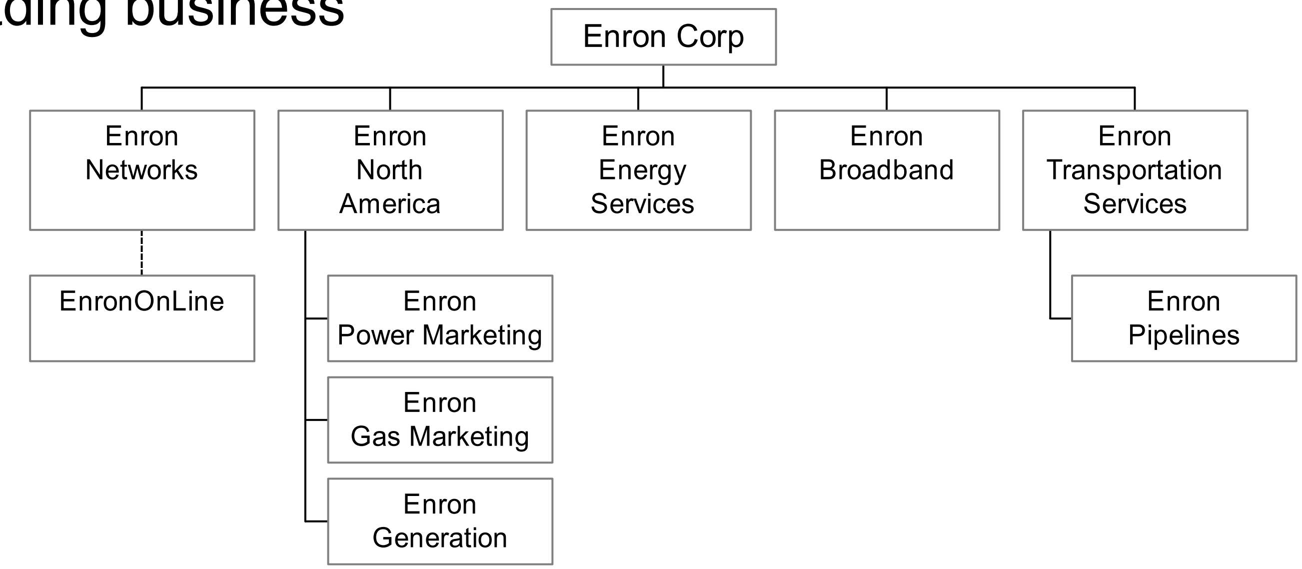


- Links may consist of
  - Relationships (e.g., friends, family, co-workers, co-authors)
  - Communications
- What can we learn about this network strictly from these connections?

Sandia National Laboratories

Enron Corp.

- U.S. corporation involved with creating energy markets
  - 7th largest by revenue
- EnronOnline: e-trading business
  - natural gas
  - electric power



- Investigations
  - U.S. Federal Energy Regulatory Commission (FERC)
    - energy market manipulation
    - involved energy traders
  - U.S. Securities and Exhchange Commission (SEC)
    - accounting fraud
    - insider trading

Sandia National Laboratories

Enron Email Data

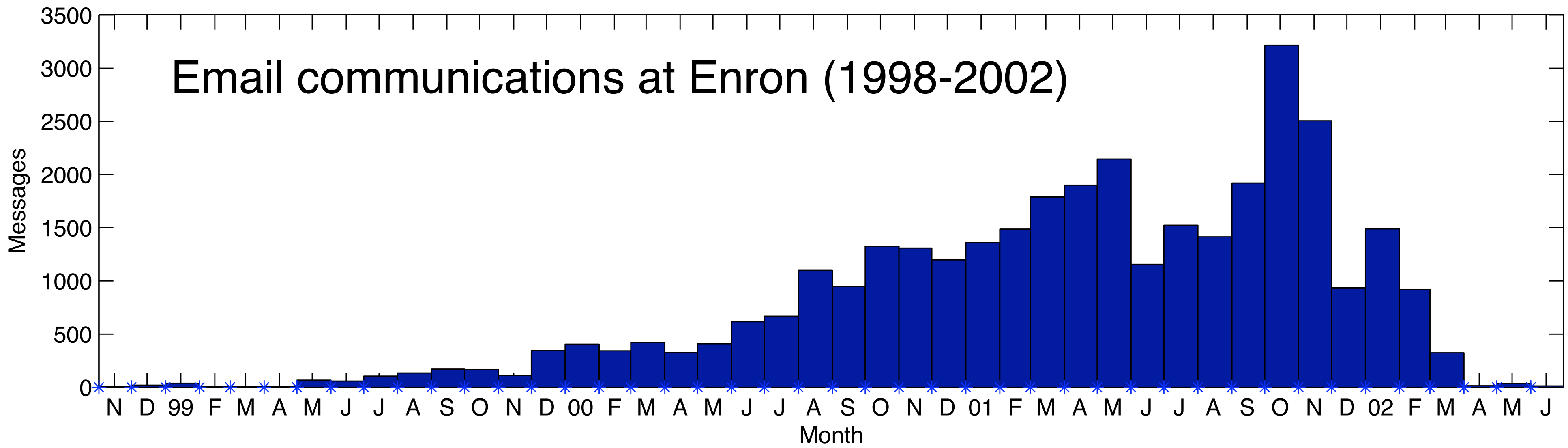
- FERC collected email of ~150 employees
  - Included emails saved in inbox, sent items, deleted items, and all other folders

- Released to the public in 2002 by FERC as part of their investigation
  - To/from, date, subject, body
  - Attachments and some names/emails removed
  - 500,000 messages

- Research uses:
  - Email classification
  - Natural language processing
  - Organizational theory/behavior
  - Social network analysis

Temporal Social Network Analysis

We use smaller dataset prepared by Priebe et al.
34,427 emails among 184 employees over 44 months



- Email folders collected at one point in time.
- Shape of histogram depends on:
  - How far back employees kept emails
  - Employment history of individual

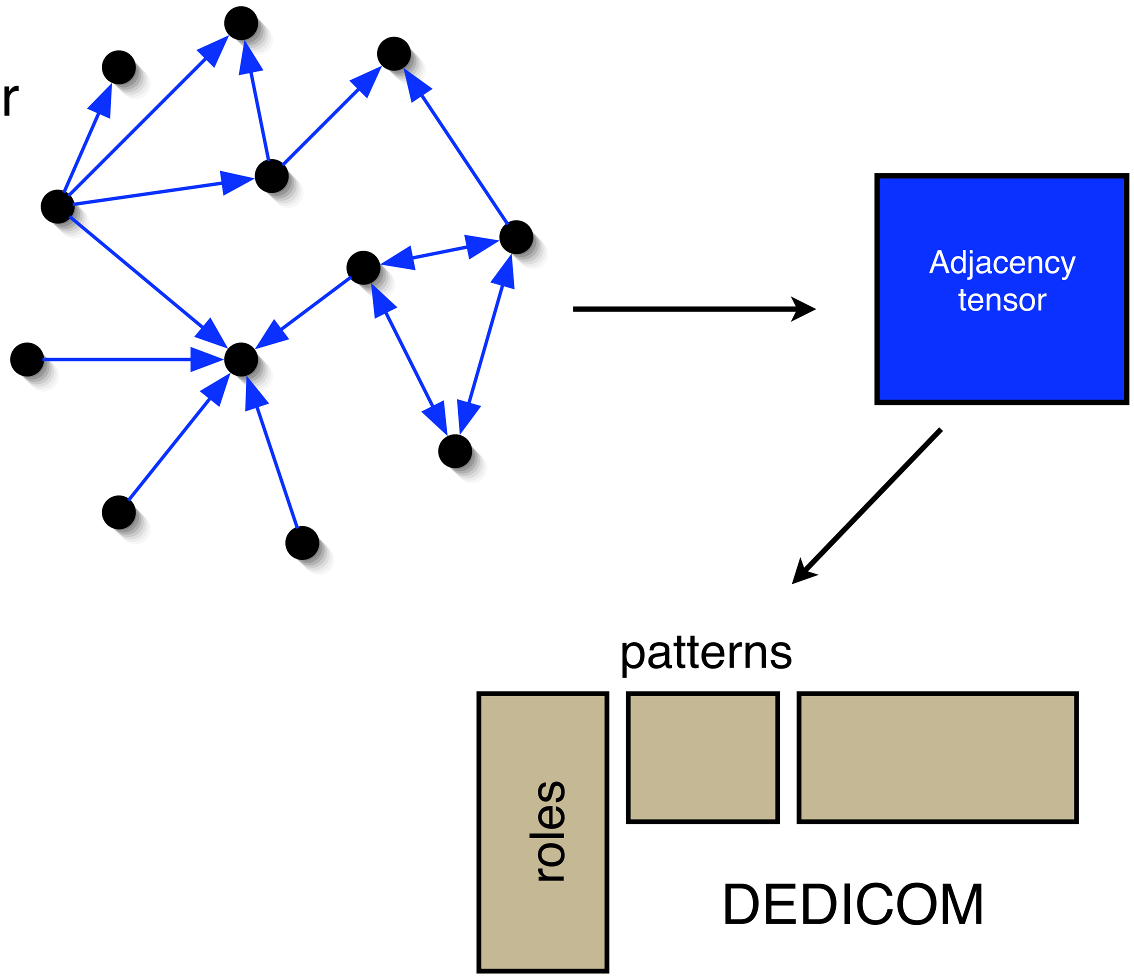
DEDICOM Experiment

- Time series of communication graphs
- Sparse tensor of size 184 x 184 x 44 (9838 nnz)
- Scaling: x number of messages scaled by log(x)+1
- Models:
  - SVD
  - 2-way DEDICOM
  - 3-way DEDICOM
  - PARAFAC

Plot of convergence history

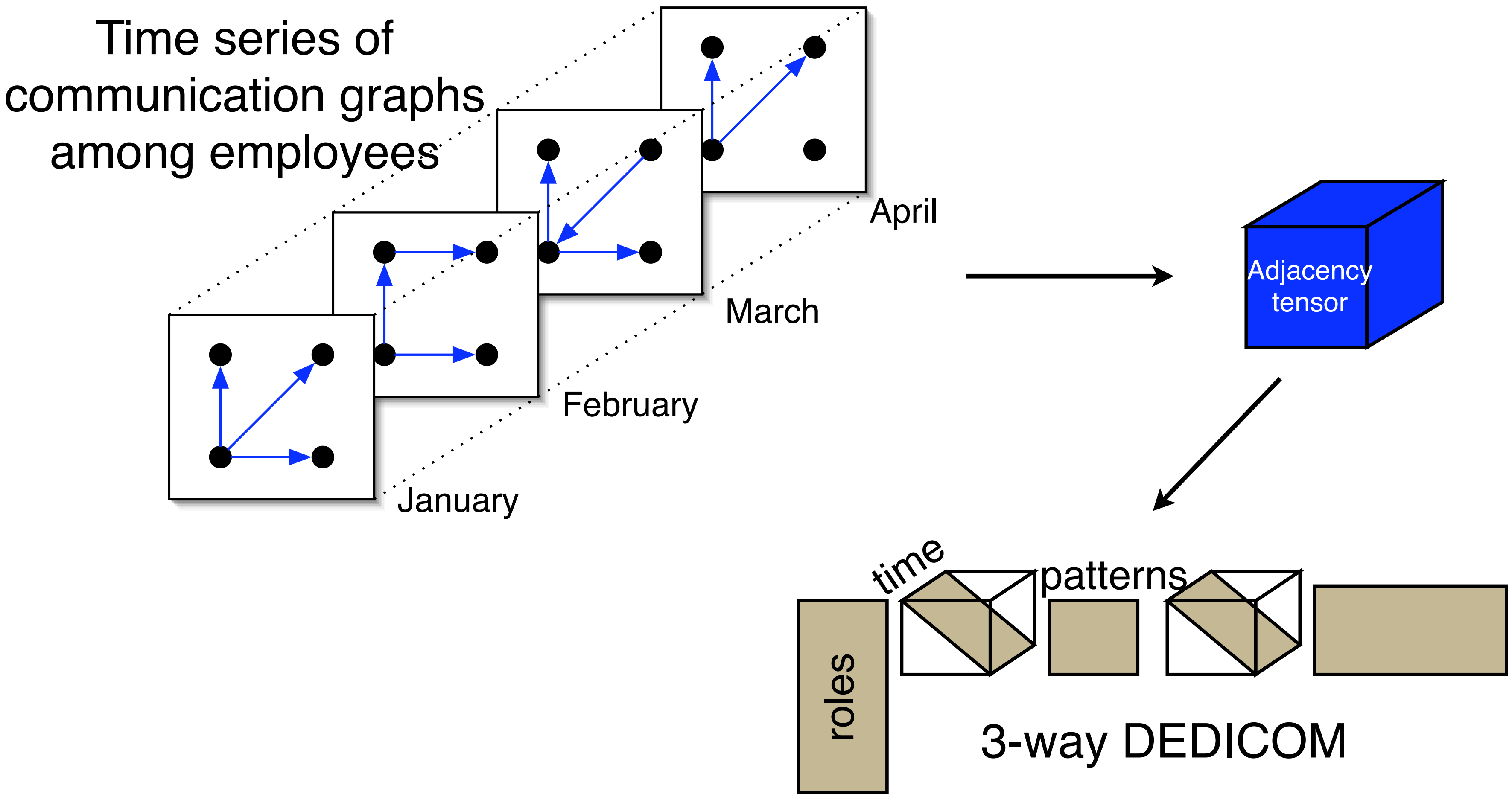
Social Network Analysis

Communication graph among employees over all times



- Description of employees by their roles
- Aggregate communication patterns among roles

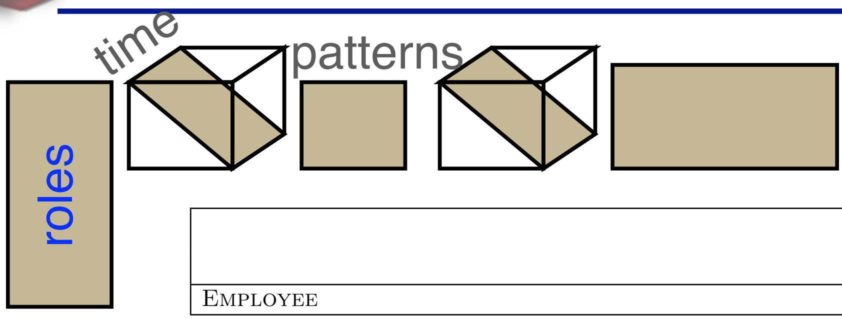
Temporal Social Network Analysis



- Unique description of employees by their roles
- Aggregate communication patterns among roles
- Behavior over time

Sandia National Laboratories

Roles of Employees



Legal | Gov't affairs | Trade execs | Pipeline

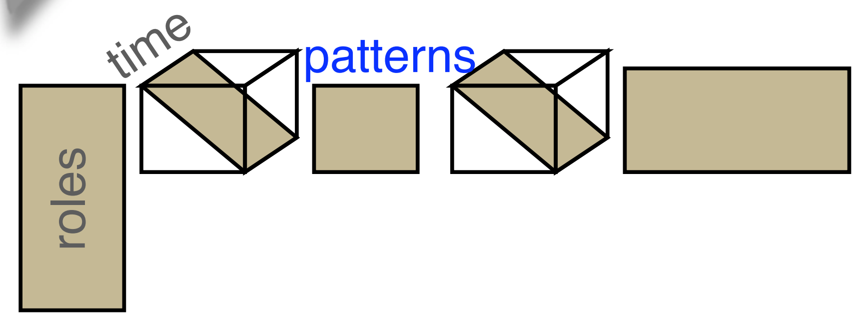
Employee	4-Dimensional Solution 1	2	3	4
T. Jones - Employee, Financial Trading Group (ENA Legal)	**0.64**	-0.01	0.02	-0.00
S. Shackleton - Employee, ENA Legal	**0.45**	-0.00	-0.01	-0.00
M. Taylor - Manager, Financial Trading Group ENA Legal	**0.37**	0.01	0.02	-0.00
S. Bailey - Legal Assistant, ENA Legal	**0.26**	-0.00	-0.01	-0.00
S. Panus - Senior Legal Specialist, ENA Legal	**0.26**	-0.00	-0.00	-0.00
M. Heard - Senior Legal Specialist, ENA Legal	**0.23**	-0.00	0.00	-0.00
J. Hodge - Asst General Counsel, ENA Legal	**0.13**	0.03	0.01	-0.00
L. Kitchen - President, Enron Online	**0.11**	**-0.09**	**0.53**	0.00
S. Dickson - Employee, ENA Legal	**0.09**	-0.00	0.00	-0.00
E. Sager - VP and Asst Legal Counsel, ENA Legal	**0.08**	0.02	**0.07**	-0.00
J. Dasovich - Employee, Government Relationship Executive	-0.01	**0.58**	0.06	0.01
J. Steffes - VP, Government Affairs	0.00	**0.53**	**-0.06**	-0.01
R. Shapiro - VP, Regulatory Affairs	-0.00	**0.40**	**0.10**	-0.00
S. Kean - VP, Chief of Staff	-0.00	**0.37**	-0.04	-0.00
R. Sanders - VP, Enron Wholesale Services	0.03	**0.16**	-0.01	-0.00
D. Delainey - CEO, ENA and Enron Energy Services	0.01	**0.09**	**0.09**	-0.00
S. Corman - VP, Regulatory Affairs	-0.00	**0.08**	-0.00	**0.20**
M. Carson - Employee, Corporate and Environmental Policy	-0.00	**0.08**	-0.02	-0.00
S. Scott - Employee, Transwestern Pipeline Company (ETS)	-0.00	**0.08**	-0.00	0.04
J. Lavorato - CEO, Enron America	0.02	-0.04	**0.49**	0.00
M. Grigsby - Director, West Desk Gas Trading	0.00	-0.03	**0.20**	-0.00
G. Whalley - President,	0.01	-0.01	**0.19**	0.00
J. Steffes - VP, Government Affairs	0.00	-0.02	**0.18**	0.00
K. Presto - VP, East Power Trading	0.01	-0.05	**0.18**	0.00
S. Beck - COO,	0.01	-0.03	**0.17**	0.00
B. Tycholiz - VP, Marketing	0.01	-0.02	**0.16**	0.00
J. Arnold - VP, Financial Enron Online	0.03	-0.04	**0.16**	-0.00
J. Williamson - Executive Assistant,	0.00	-0.02	**0.14**	0.01
K. Watson - Employee, Transwestern Pipeline Company (ETS)	-0.00	-0.00	0.01	**0.59**
M. Lokay - Admin. Asst., Transwestern Pipeline Company (ETS)	-0.00	0.01	0.01	**0.42**
L. Donoho - Employee, Transwestern Pipeline Company (ETS)	-0.00	0.01	0.01	**0.35**
M. McConnell - Employee, Transwestern Pipeline Company (ETS)	0.00	-0.00	0.01	**0.26**
L. Blair - Employee, Northern Natural Gas Pipeline (ETS)	-0.00	0.00	0.00	**0.22**
K. Hyatt - Director, Asset Development TW Pipeline Business (ETS)	-0.00	0.01	0.00	**0.20**
D. Schoolcraft - Employee, Gas Control (ETS)	-0.00	0.00	0.00	**0.18**
T. Geaccone - Manager, (ETS)	0.00	-0.00	0.01	**0.17**
R. Hayslett - VP, Also CFO and Treasurer	0.00	-0.00	0.02	**0.16**

Groups (left of table): Legal; Gov't affairs; Execs - trading; Pipeline employees

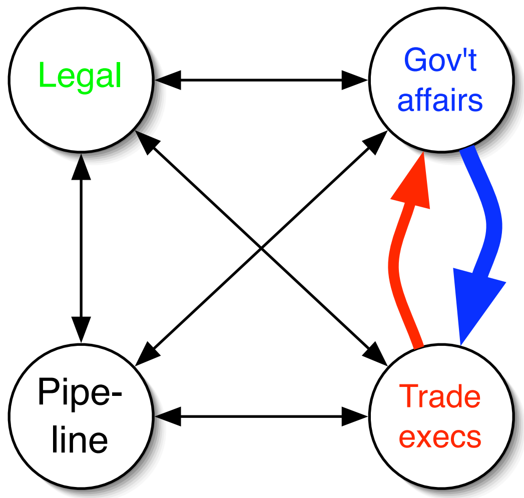
Identify shared characteristics to label group

Sandia National Laboratories

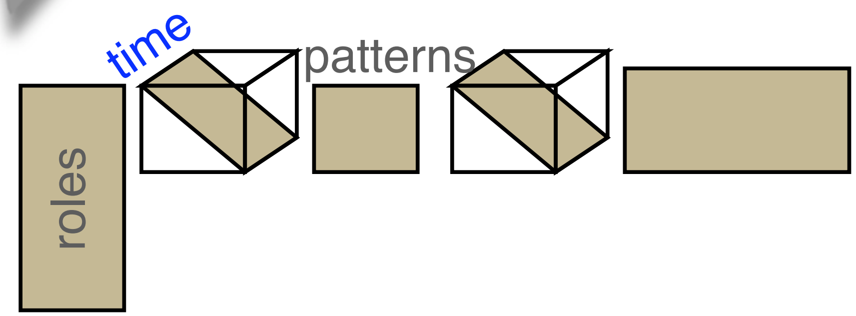
Communication Patterns

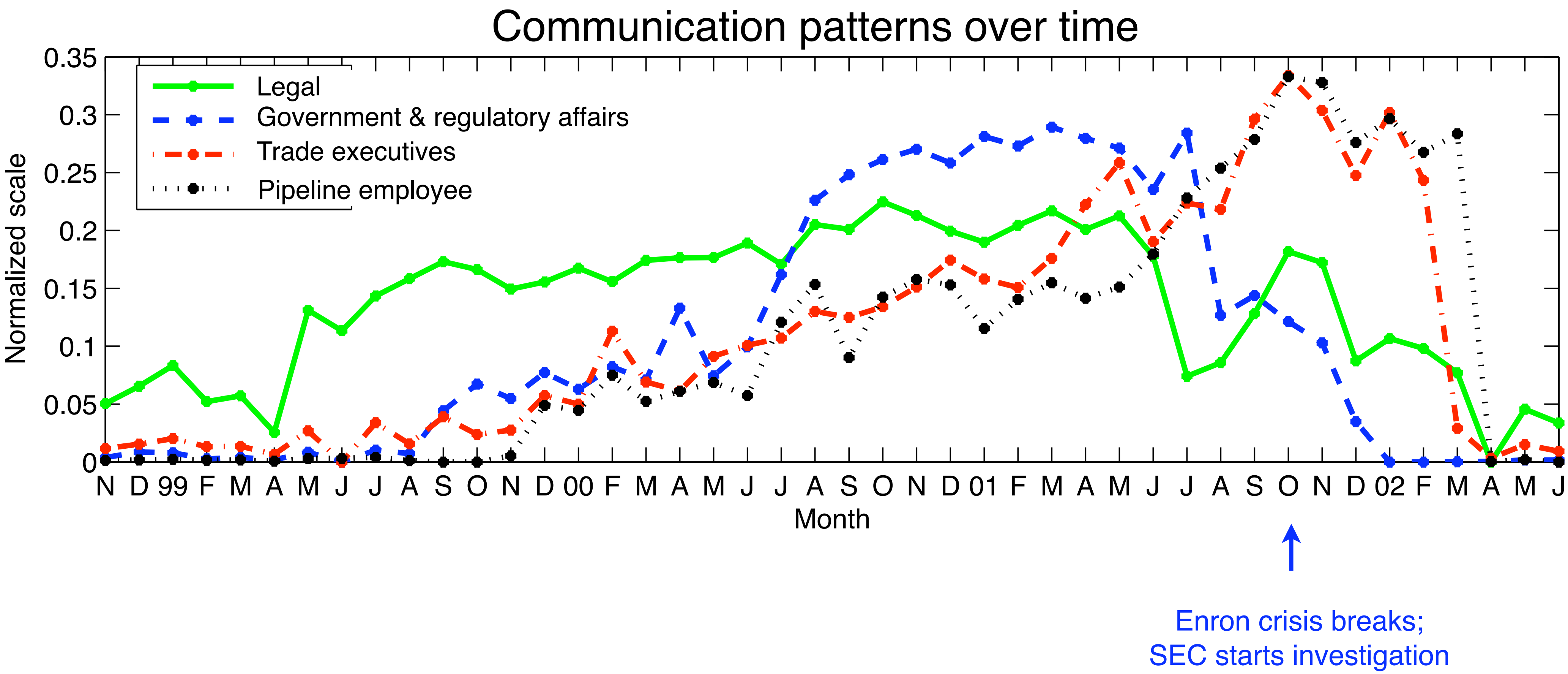


	Legal	Gov't affairs	Trade execs	Pipeline
Legal	440.2	1.6	-15.0	0.4
Government & regulatory affairs	1.6	278.3	135.4	1.6
Trade executives	-29.3	70.7	201.6	-6.2
Pipeline employees	1.4	-4.6	-7.5	172.3



Temporal Patterns





Summary

- Improvements to DEDICOM
  - New procedure for finding A
  - Newton step for finding D
- Modifications to handle sparse data arrays
  - Least squares problem
  - Compression
- Novel approach to social network analysis using DEDICOM
  - Roles of employees
  - Communication patterns among roles and over time
- Many future research directions!
  - Constrained DEDICOM
  - Nonnegative DEDICOM

More Information

bwbader@sandia.gov
http://www.cs.sandia.gov/~bwbader/

- Tensor Classes:
  - Tech report SAND2004-5189 available on website
  - Paper to appear in ACM Trans. Math. Softw.
  - (sparse_tensor to be released soon)

- DEDICOM paper:
  - Tech report SAND2006-2161 available on website

Sandia National Laboratories