SAND2007-7044C

# Temporal Analysis of Semantic Graphs using ASALSAN

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

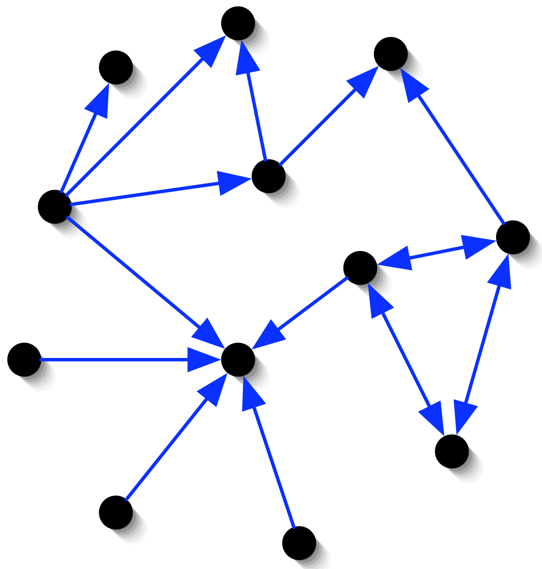
International Conference on Data Mining
October 31, 2007

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.

# Common Graph Analysis Technique

For example:

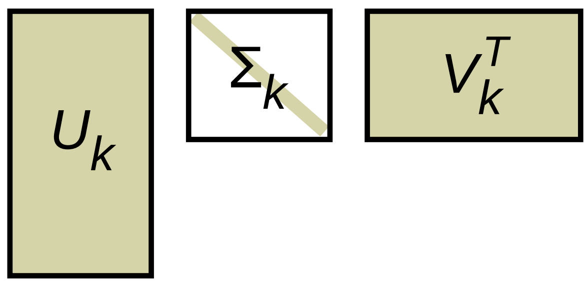
Web search - HITS (Kleinberg, 1998)



Adjacency matrix

Truncated SVD

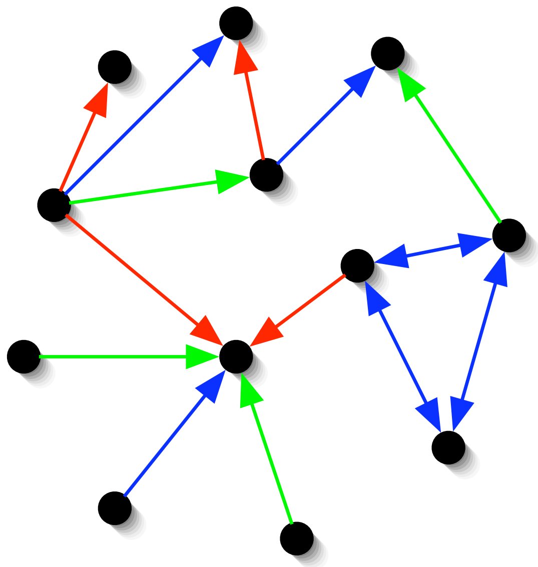
$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$




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

But we may have ignored critical information by not considering edge metadata!

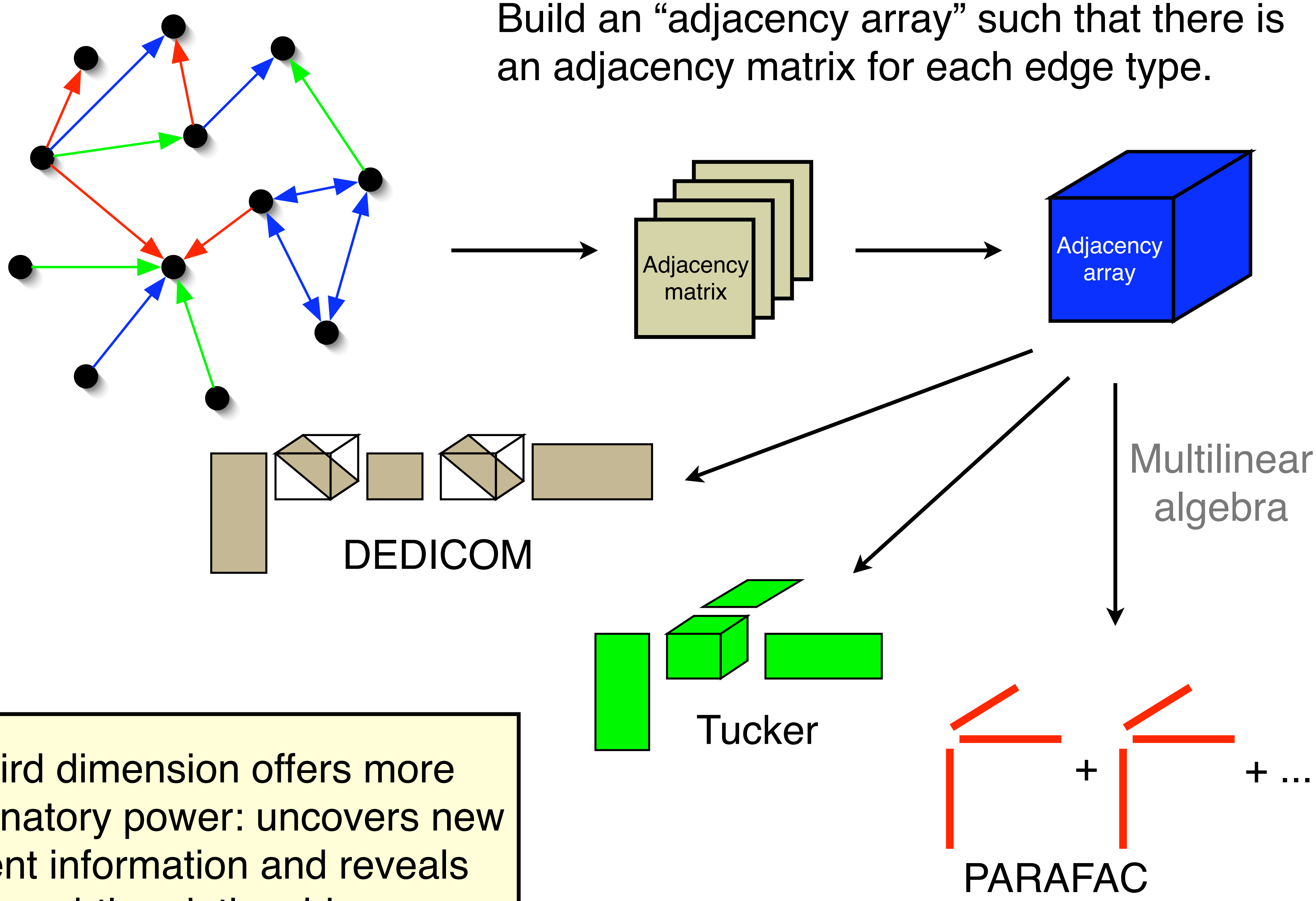
# Semantic Graphs



- Different types of edges
- Examples
  - WWW (anchor text)
  - Subway map
  - Email communications (time stamp, to/cc)

# New Paradigm: "Multidimensional Data Mining"

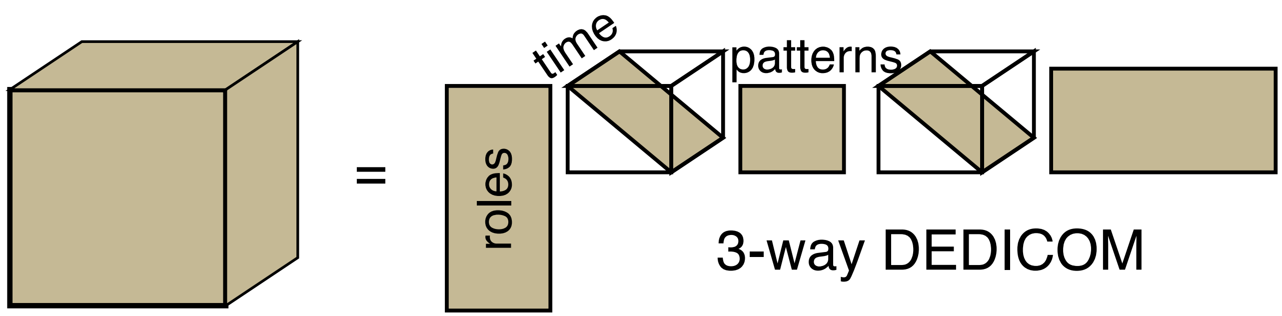
Build an "adjacency array" such that there is an adjacency matrix for each edge type.

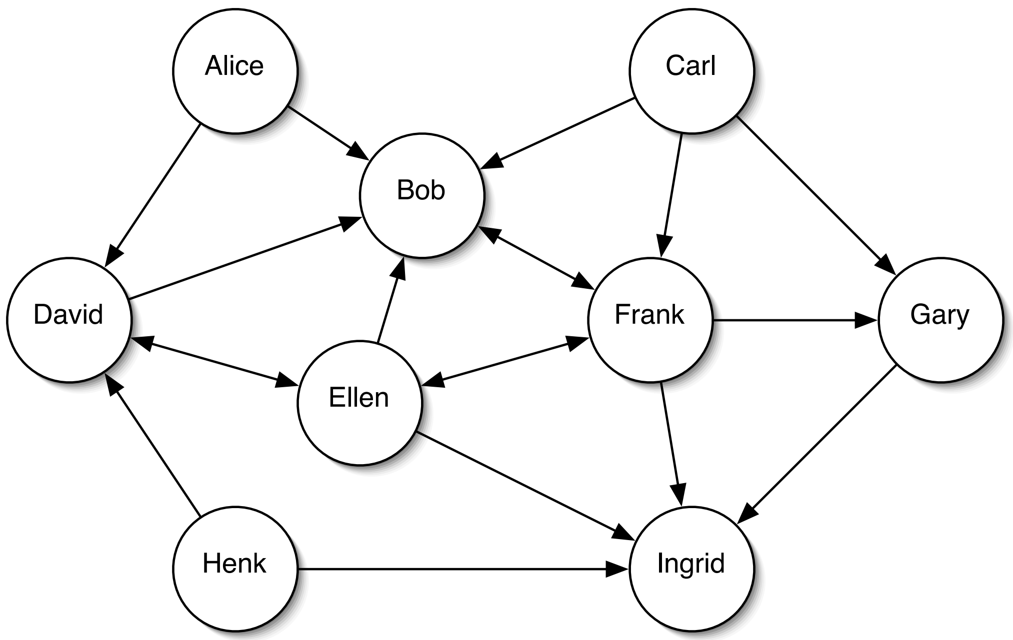



Third dimension offers more explanatory power: uncovers new latent information and reveals subtle relationships

# Objective

Use ASALSAN to fit DEDICOM model to analyze a semantic graph of timestamp-labeled edges




Sandia National Laboratories

# DEDICOM

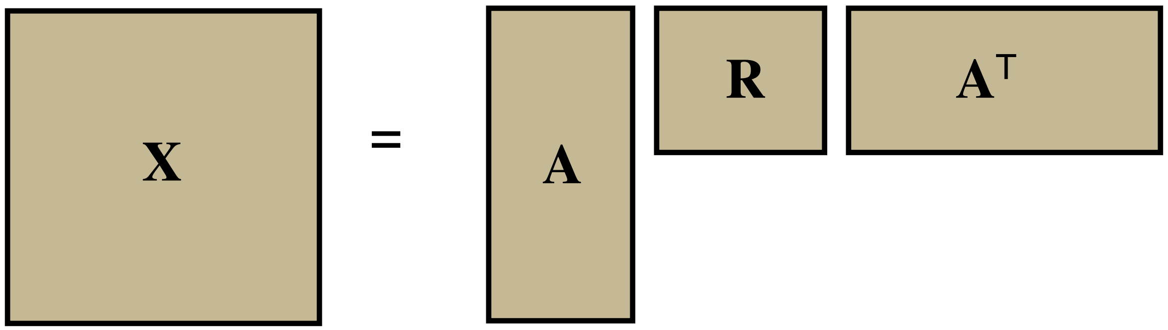
- DEcomposition into DIrectional COMponents

- Introduced in 1978 by Harshman

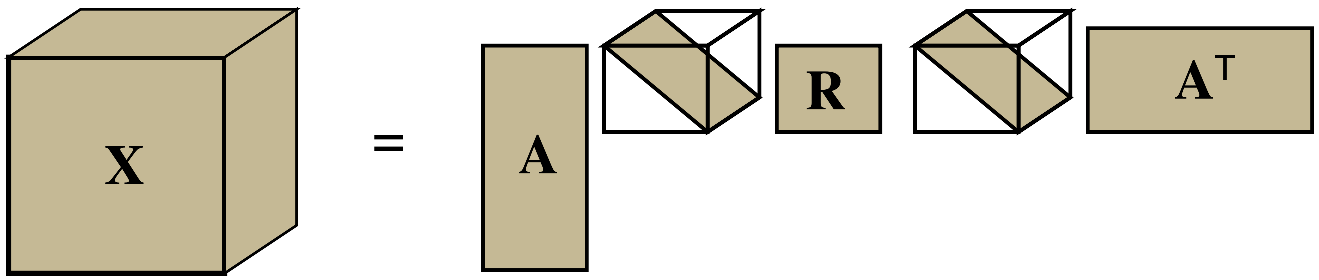
- Past applications
  - Study asymmetries in telephone calls among cities
  - Marketing research
    - car switching: car owners and what they buy next
    - free associations of words
      - words to describe hair in advertising shampoo: “body” evokes “fullness” more often than “fullness” evokes “body”
  - Asymmetric measures of world trade (import/export)

- Variations
  - Three-way DEDICOM
  - Constrained DEDICOM

# DEDICOM Models & Algorithms




- Generalized Takane method (Takane, 1985; Kiers et al., 1990)
- New algorithm

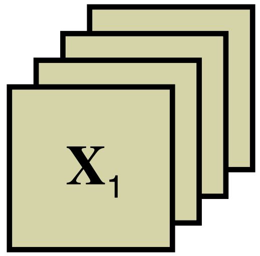



- Kiers' method (Kiers, 1993)
- New algorithm

All are "alternating" algorithms

# 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

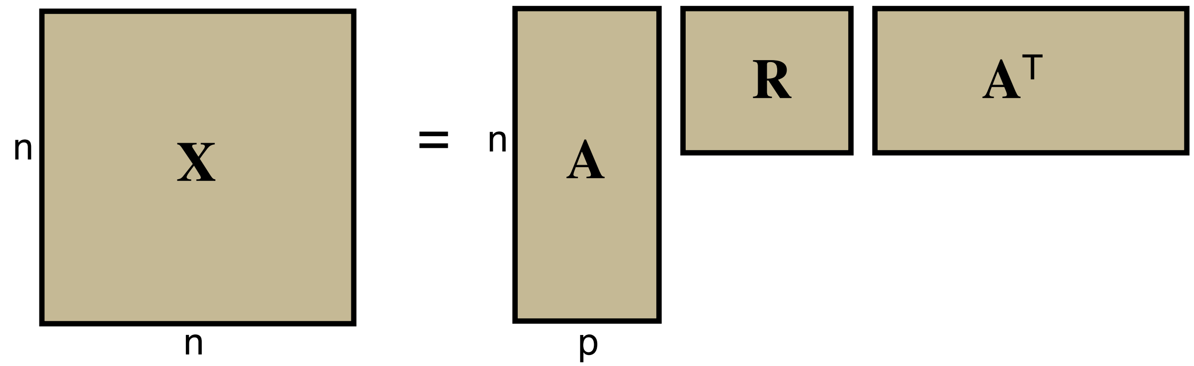
$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

  - Hadamard product (elementwise)

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

# Two-way DEDICOM

Single domain model




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

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

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

s.t. $\mathbf{A}$ orthogonal

- $\mathbf{A}$ ($n$ x $p$) is an orthogonal matrix of loadings or weights
- $\mathbf{R}$ ($p$ x $p$) is a dense matrix that captures asymmetric relationships

- Decomposition is not unique
  - $\mathbf{A}$ can be transformed with no loss of fit to the data
  - Nonsingular transformation $\mathbf{Q}$:

$$\mathbf{A}\mathbf{R}\mathbf{A}^T = (\mathbf{A}\mathbf{Q})(\mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-T})(\mathbf{A}\mathbf{Q})^T$$

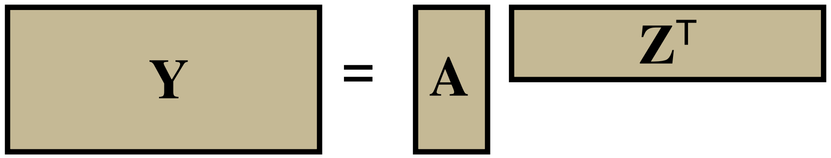
  - Usually “fix” $\mathbf{A}$ with some standard rotation (e.g., VARIMAX)

# New Algorithm

## Solving for $\mathbf{A}$:

Stack data and model "side by side" in a single equation

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




...and solve least-squares problem: $\min_{\mathbf{A}} \left\| \mathbf{Y} - \mathbf{A}\mathbf{Z}^T \right\|_F^2$

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

or

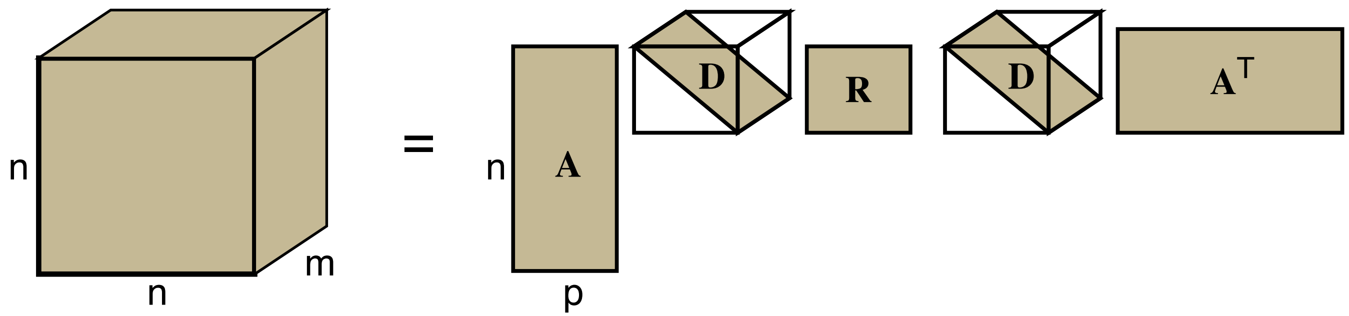
$$\mathbf{A}_{new} = \left(\mathbf{X}\mathbf{A}\mathbf{R}^T + \mathbf{X}^T\mathbf{A}\mathbf{R}\right)\left(\mathbf{R}(\mathbf{A}^T\mathbf{A})\mathbf{R}^T + \mathbf{R}^T(\mathbf{A}^T\mathbf{A})\mathbf{R}\right)^{-1}.$$

## Solving for $\mathbf{R}$:

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

Sandia National Laboratories

# 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 a matrix of loadings or weights (not necessarily orthogonal)
- $\mathbf{R}$ ($p$ x $p$) is a dense matrix that captures asymmetric relationships
- $\mathbf{D}$ ($p$ x $p$ x $m$) is a tensor with diagonal frontal slices giving the weights of the columns of $\mathbf{A}$ for each slice in third mode

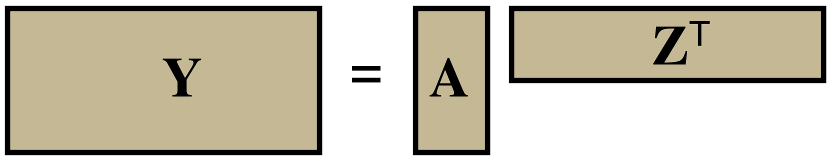
- *Unique* solution with enough slices of $\mathbf{X}$ with sufficient variation
  - i.e., no rotation of $\mathbf{A}$ possible
  - greater confidence in interpretation of results

# New Algorithm - ASALSAN

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

## Solving for $\mathbf{A}$:

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




$$\mathbf{A} = \mathbf{Y}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$$

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

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

Sandia National Laboratories

# New Algorithm - ASALSAN

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

Solving for $\mathbf{D}$:

Use Newton's method to solve the optimization problem for $d = \text{diag}(\mathbf{D}_i)$

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

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

Hessian: $h_{st} = -2\sum_{i,j} \Big[ (\mathbf{X} - \mathbf{ADRDA}^T) * (\mathbf{a}_s r_{st}\mathbf{a}_t^T + \mathbf{a}_t r_{ts}\mathbf{a}_s^T)$

$$- (\mathbf{AD}\mathbf{r}_s\mathbf{a}_s^T + \mathbf{a}_s\mathbf{r}_{s:}\mathbf{DA}^T) * (\mathbf{AD}\mathbf{r}_t\mathbf{a}_t^T + \mathbf{a}_t\mathbf{r}_{t:}\mathbf{DA}^T) \Big]_{i,j}$$

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 ($p$ x $p$)

Sandia National Laboratories

# New Algorithm - ASALSAN

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

Solving for $\mathbf{R}$:

Use the approach in (Kiers, 1993)

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

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

# Algorithm Costs

Updating $\mathbf{A}$ is most expensive part

Dominant costs:

linear in nnz of $\mathbf{X}_i$:
- $\mathbf{Q}^T\mathbf{X}_i\mathbf{Q}$
- $\mathbf{X}_i\mathbf{A}\mathbf{R}^T$
- $\mathbf{X}_i^T\mathbf{A}\mathbf{R}$

$\mathcal{O}(p^2n)$:
- $\mathbf{A}^T\mathbf{A}$
- QR factorization of $\mathbf{A}$

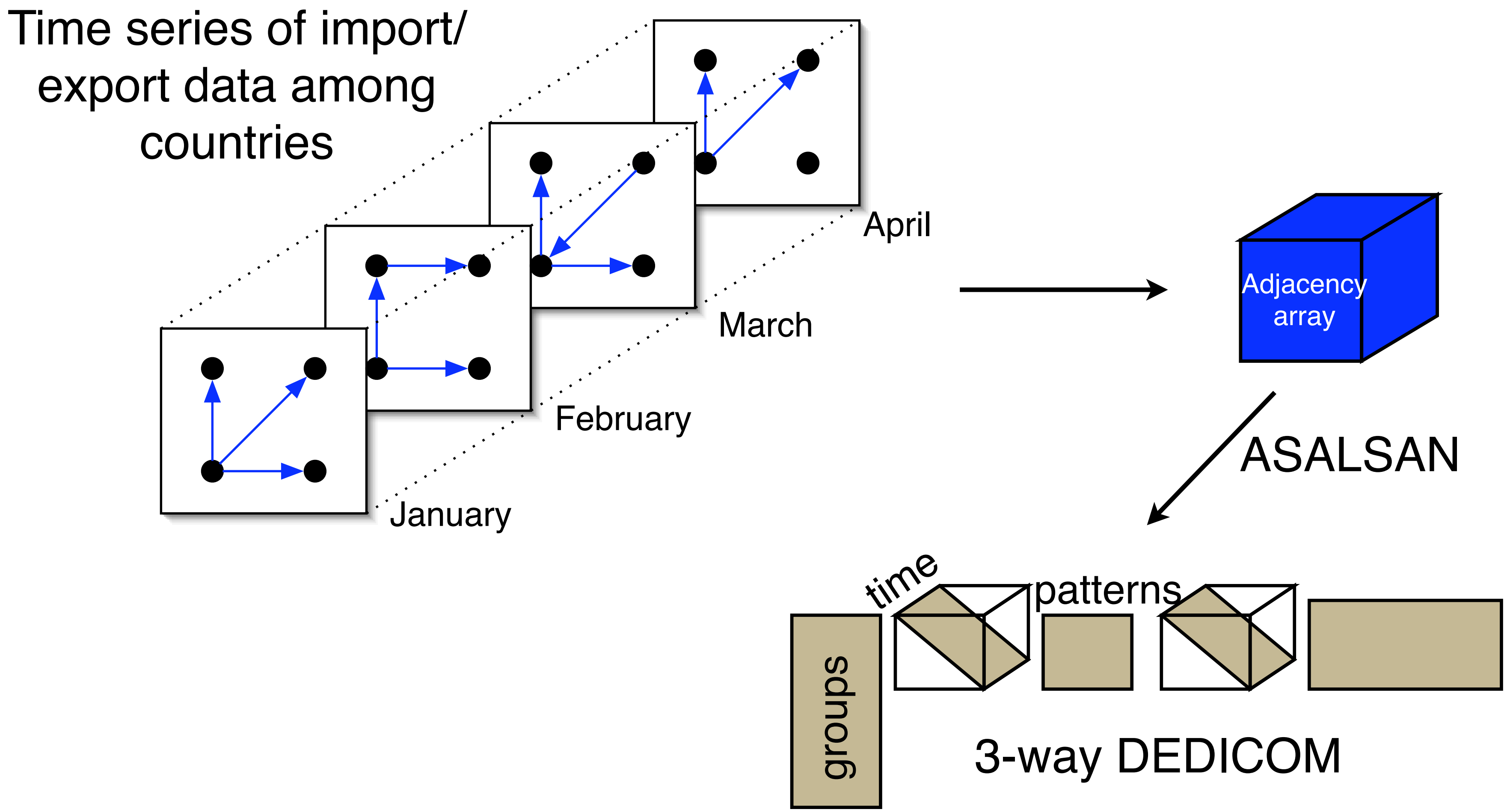
Time in seconds per iteration (avg iterations)

| Algorithm | World trade | | Enron | |
|---|---|---|---|---|
| ASALSAN | 0.069 | (50) | 0.85 | (184) |
| NN-ASALSAN | 0.083 | (47) | 1.0 | (74) |
| Kiers [23] | 0.022 | (67) | 22.3 | (400+) |

# Application: World Trade

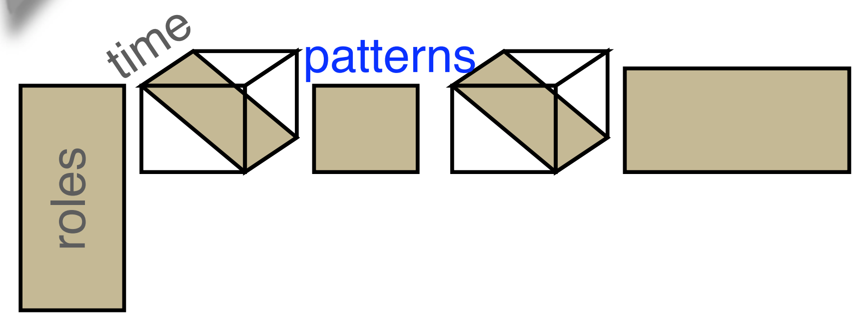
- Graph of annual import/export data
- Are there any patterns in global trade?

# Temporal World Trade Analysis

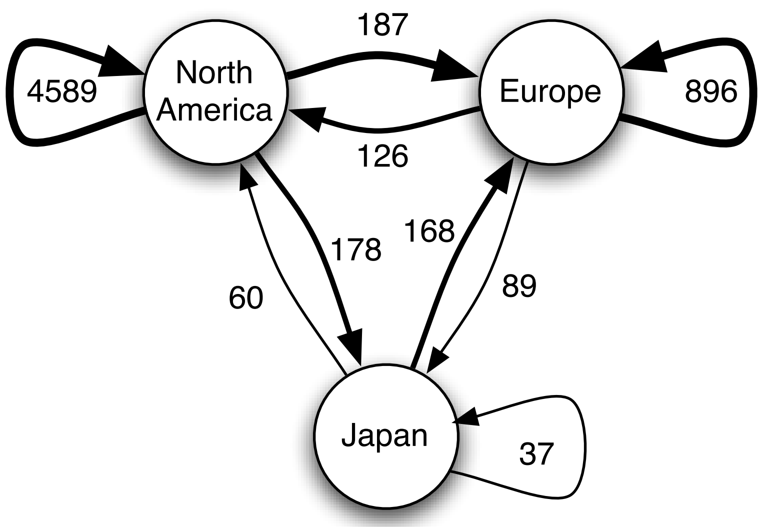



- Unique categorization of countries
- Aggregate trade patterns among regions
- Pattern over time

# World Trade Patterns



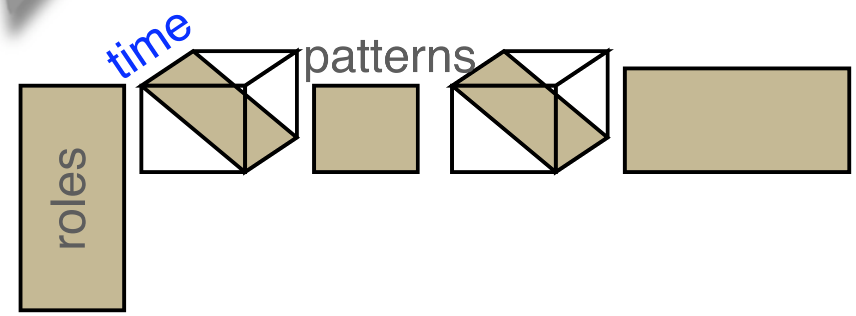

| | #1 | #2 | #3 |
|---|---|---|---|
| #1 North America | 4589 | 187 | 178 |
| #2 Europe | 126 | 896 | 89 |
| #3 Japan | 60 | 168 | 37 |

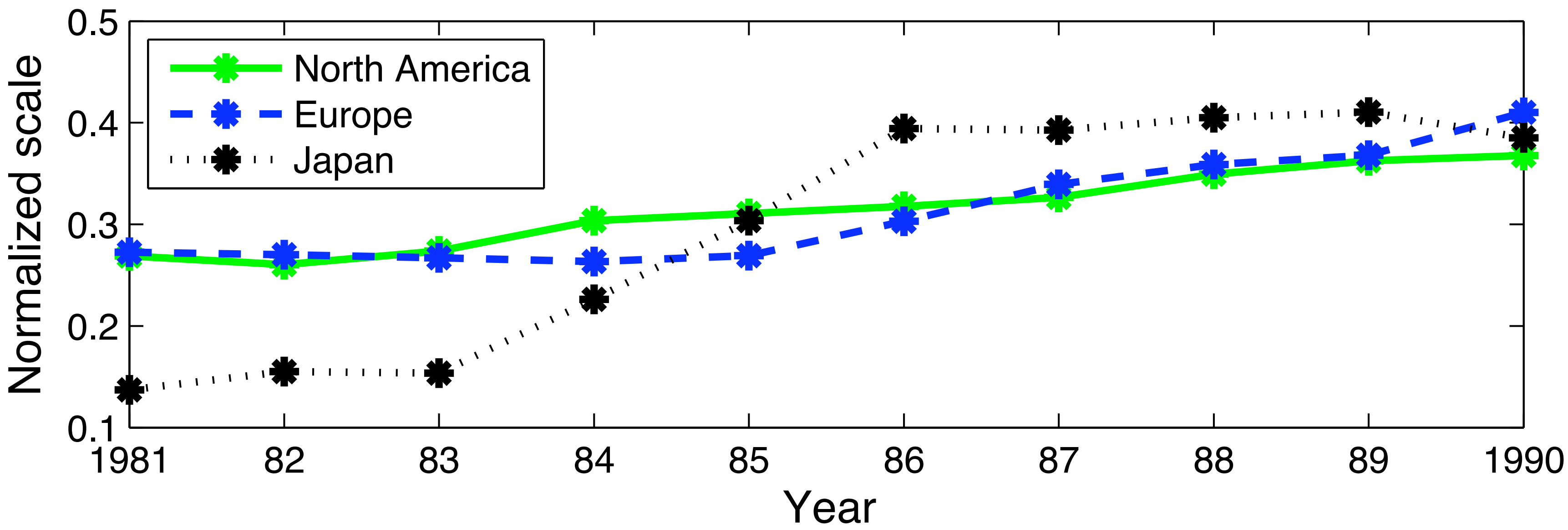



- Mostly trade within region
- Some large exchanges
- Asymmetry in exchange

Sandia National Laboratories

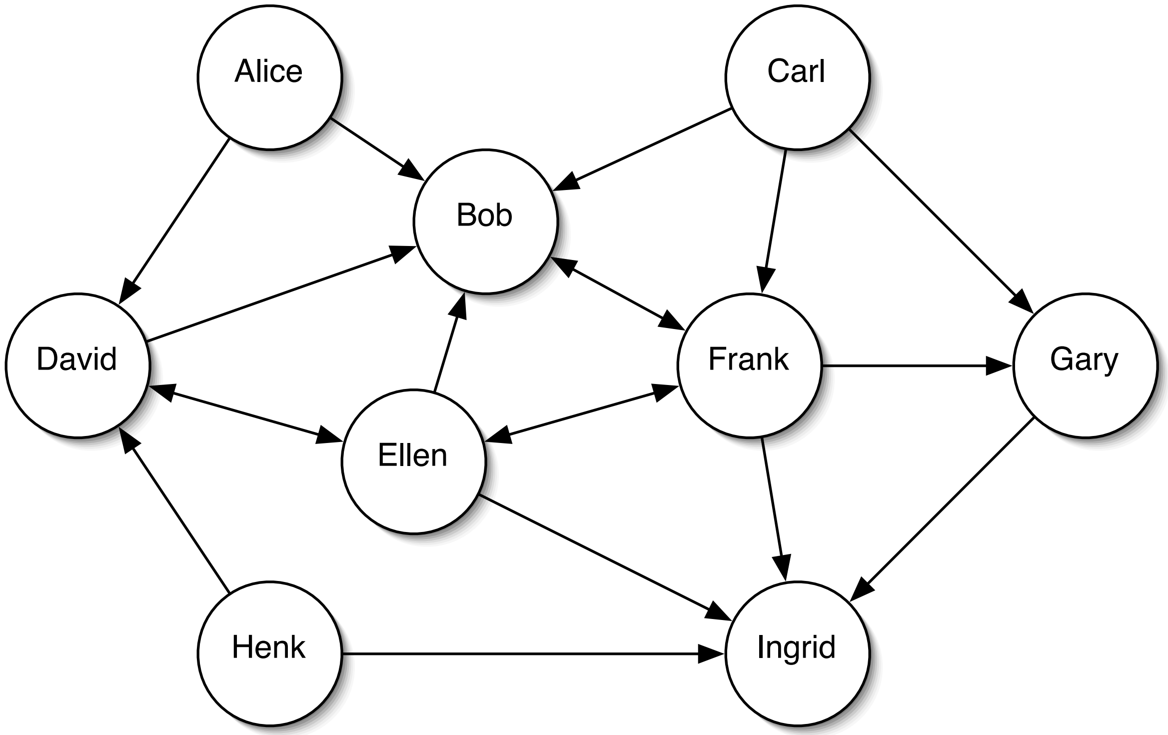
# Temporal Patterns in World Trade




Global recession in early 80's

# Application: Enron Email Analysis

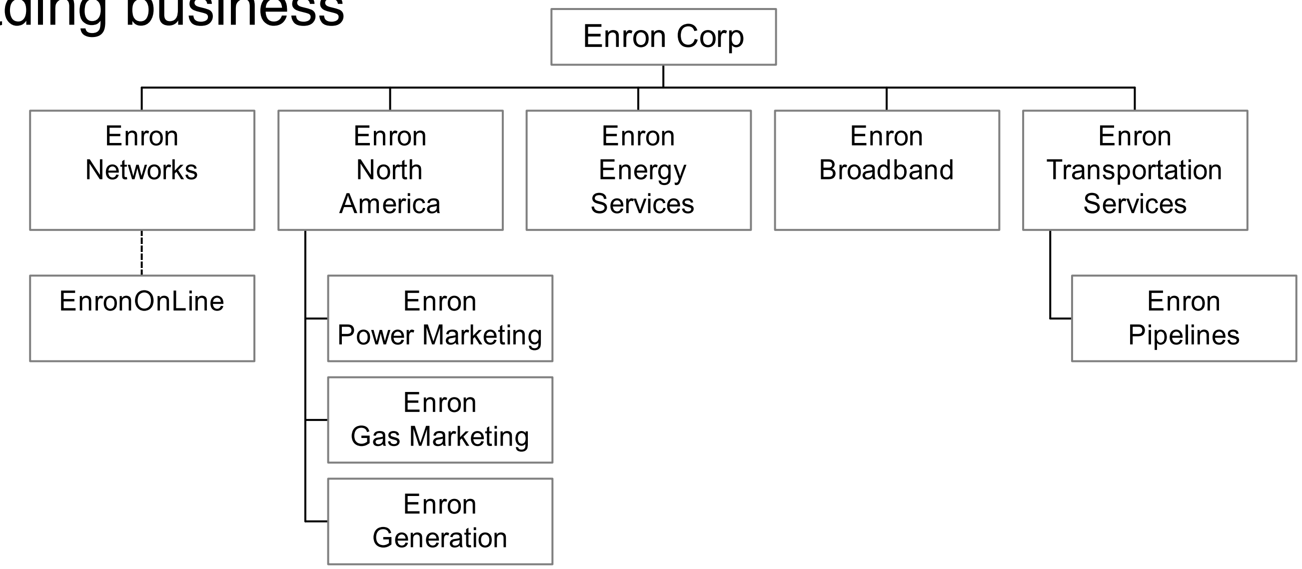



- Links consist of email communications

- What can we learn about this network strictly from their communication patterns?   (Social network analysis)

# 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 Exchange Commission (SEC)
    - accounting fraud
    - insider trading

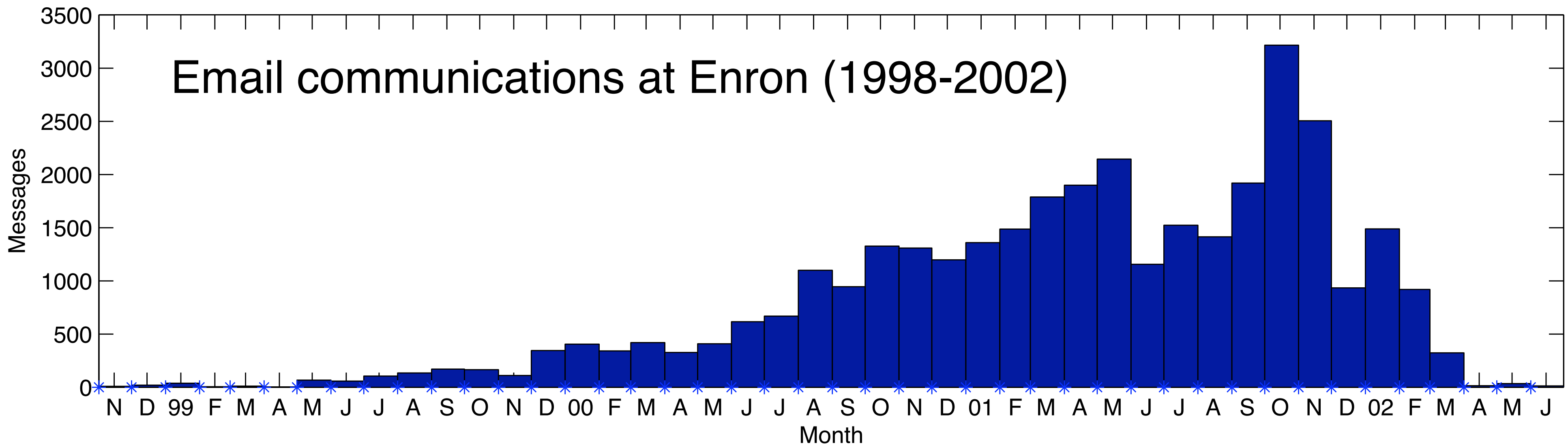
# Enron Email Data

- FERC collected email of ~150 employees as evidence
  - 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
  - Approx. 500,000 email messages

# Smaller Enron Data Set

We used a smaller data set prepared by Priebe et al.
34,427 emails among 184 employees over 44 months

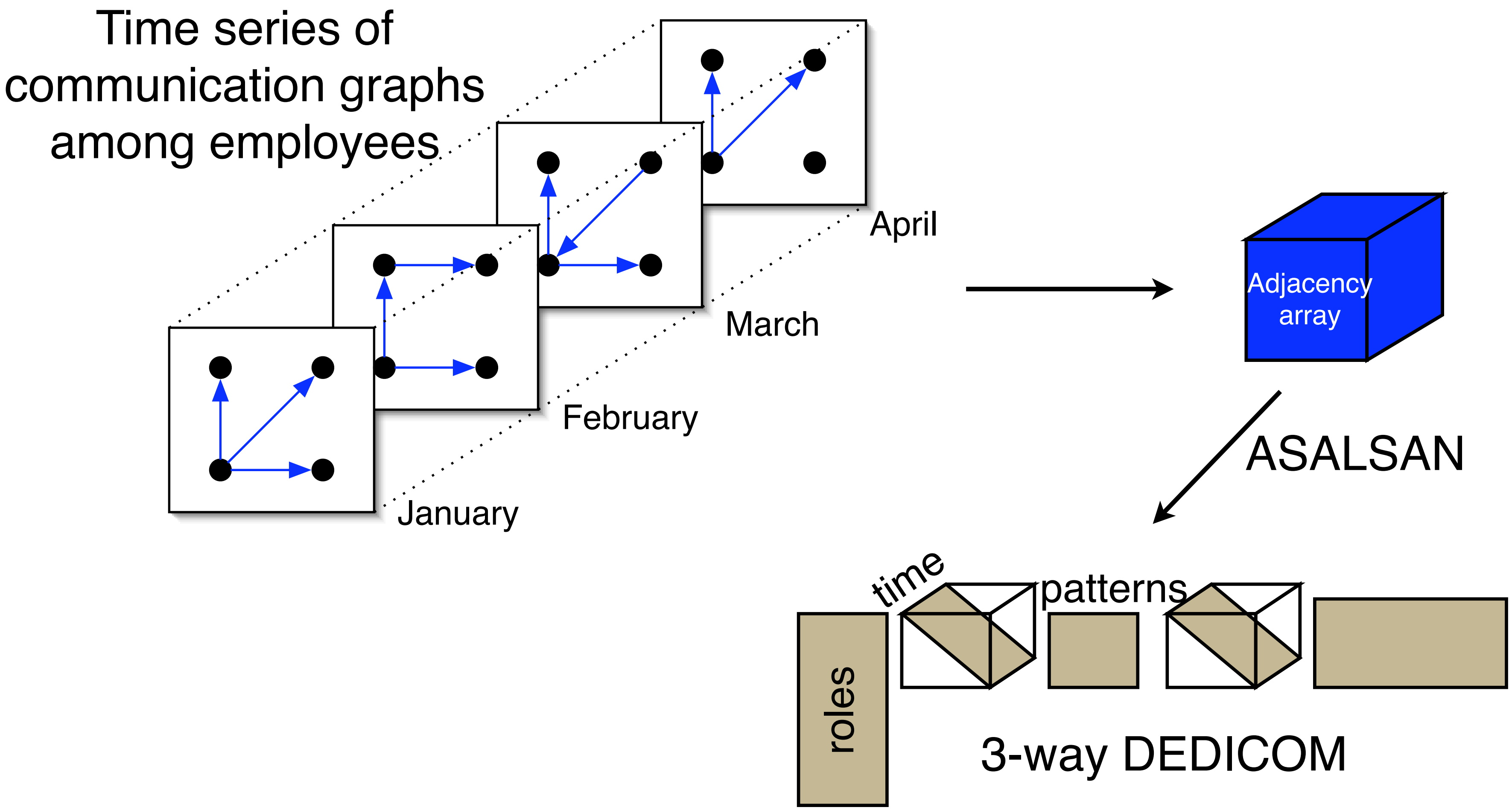



- Limited information on the 184 employees
- No org chart

# Enron Experiment

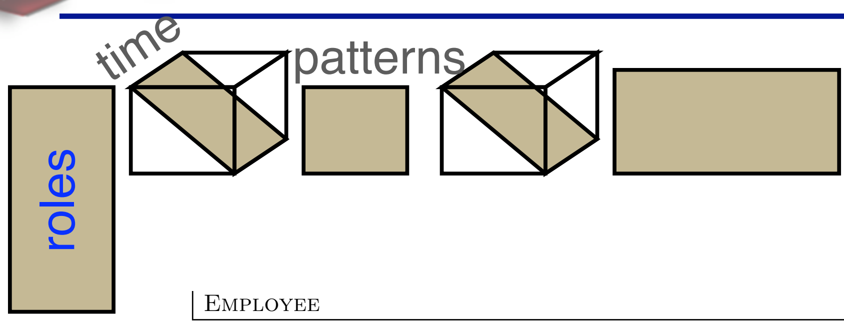
- Aggregate communications
  - Sparse matrix of size 184 x 184 (3007 nonzeros)
- Time series of communication graphs
  - Sparse tensor of size 184 x 184 x 44 (9838 nonzeros)
- Weighted adjacency matrix
  - scaling: x number of messages scaled by log(x)+1
  - other common choices give similar results
- Models:
  - SVD
  - 2-way DEDICOM
  - 3-way DEDICOM (via ASALSAN and NN-ASALSAN)

# Temporal Social Network Analysis




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

# Roles of Employees

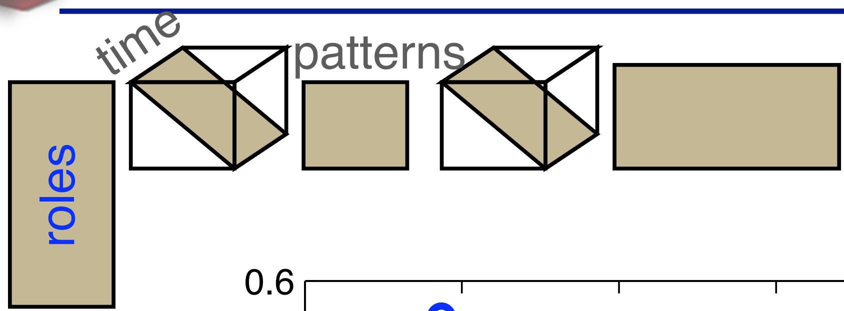


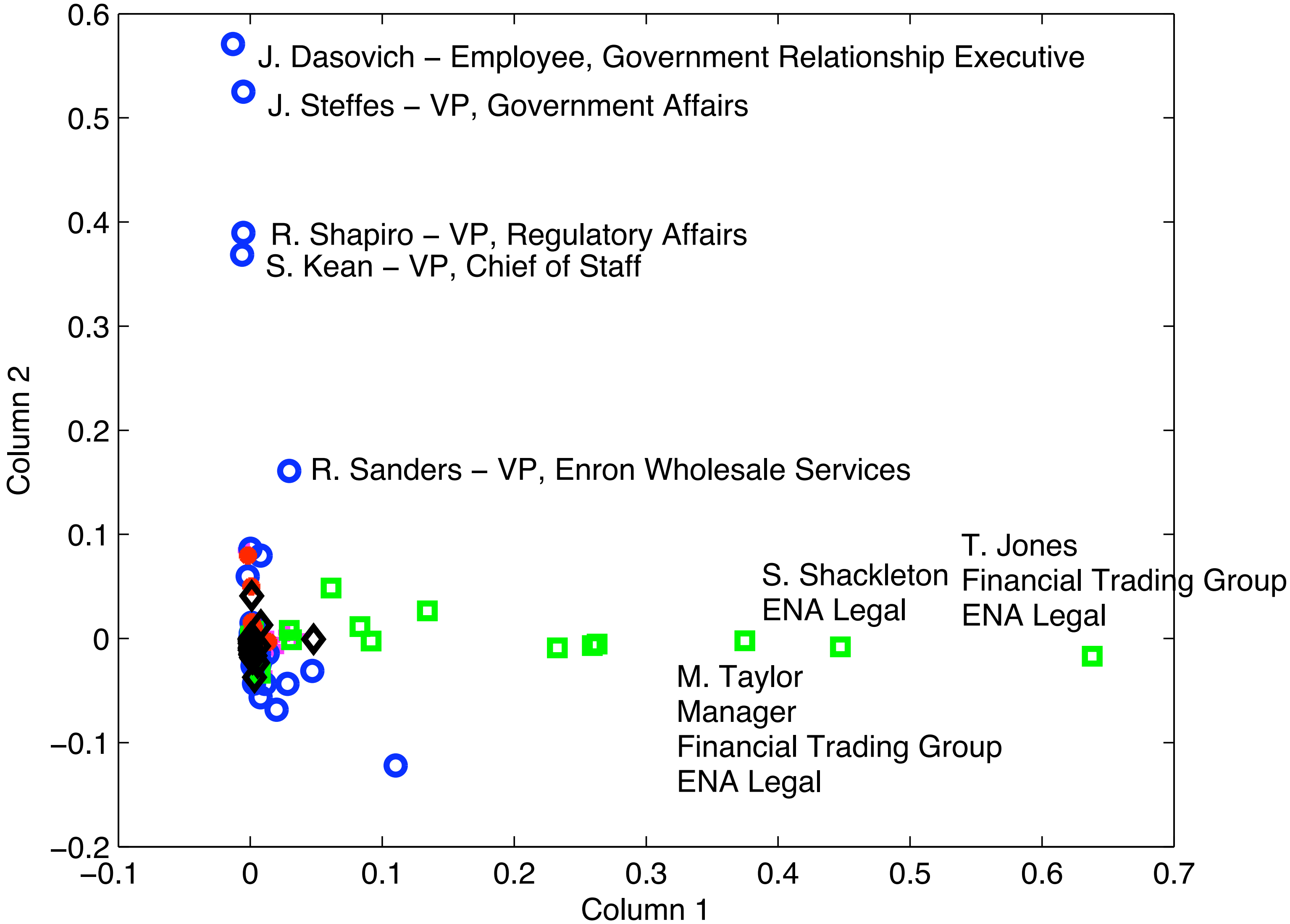

| Group | Employee | 1 (Legal) | 2 (Gov't affairs) | 3 (Trade execs) | 4 (Pipeline) |
|---|---|---|---|---|---|
| Legal | T. Jones - Employee, Financial Trading Group (ENA Legal) | **0.64** | -0.01 | 0.02 | -0.00 |
| | S. Shackleton - Employee, ENA Legal | **0.45** | -0.00 | -0.01 | -0.00 |
| | M. Taylor - 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 |
| Gov't affairs | 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 |
| Execs - trading | 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 |
| Pipeline employees | 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** |

Identify shared characteristics to label group

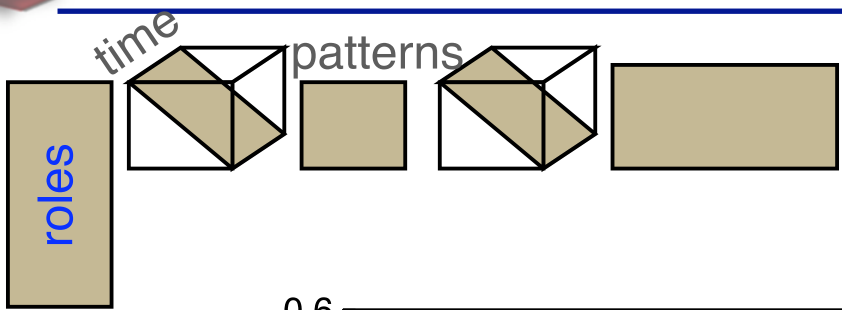
Sandia National Laboratories

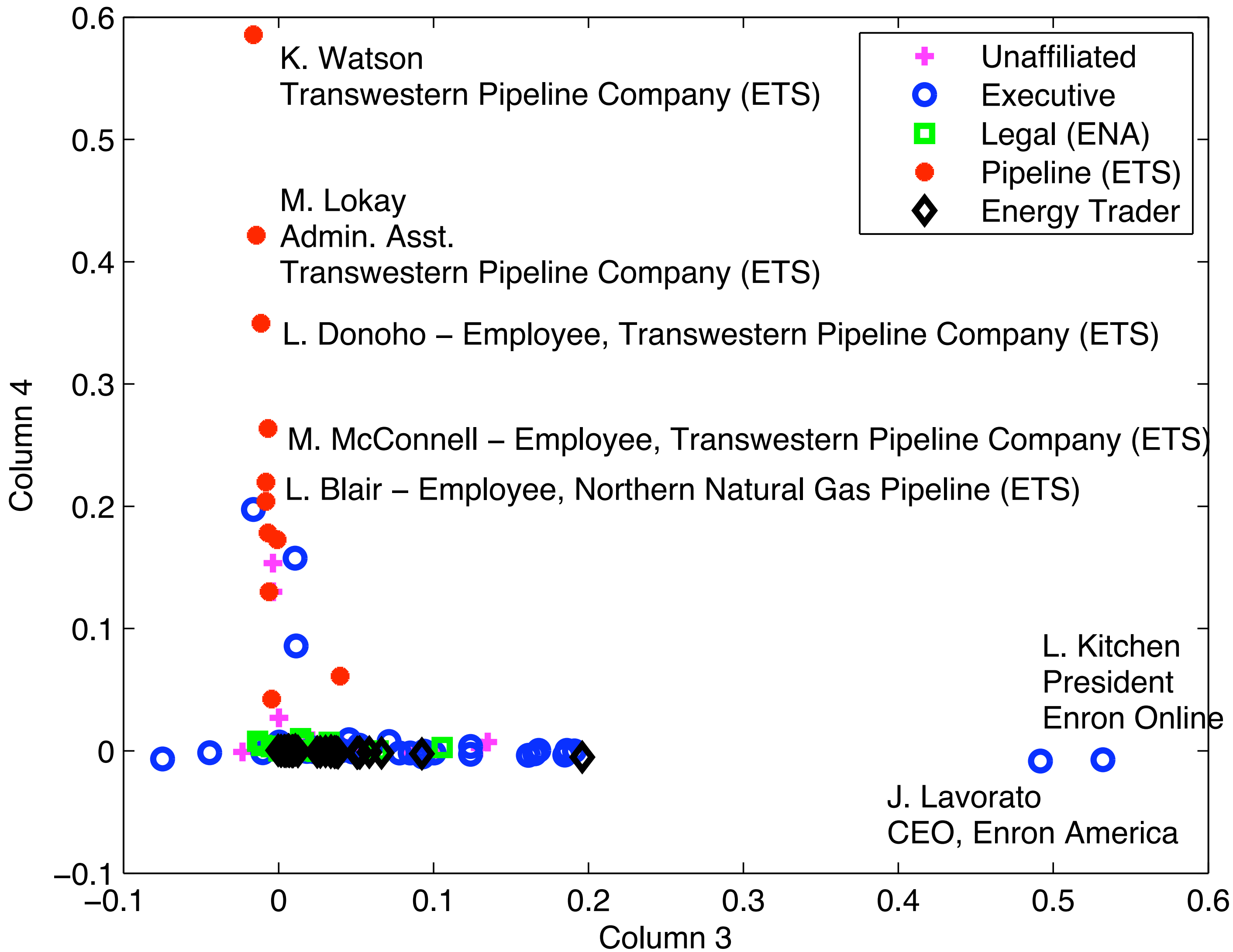
# Roles of Employees



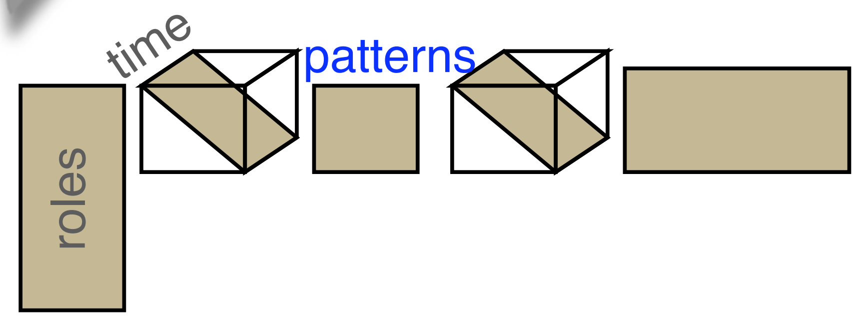
# Roles of Employees



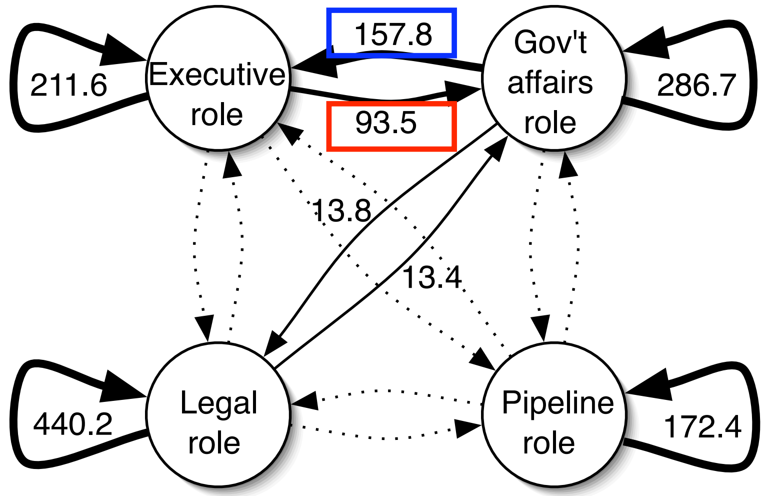




Sandia National Laboratories

# Communication Patterns

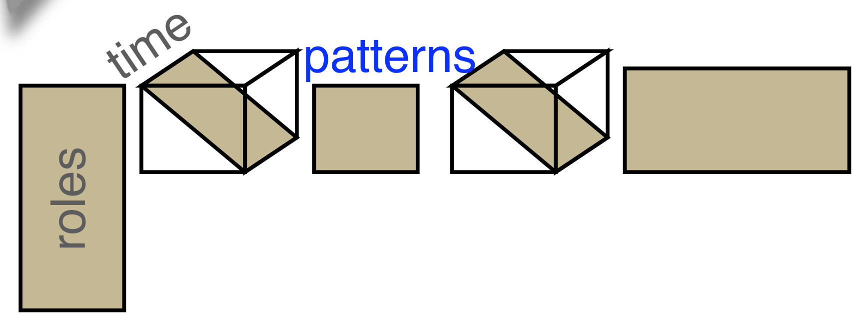



| | Legal | Gov't affairs | Trade execs | Pipeline |
|---|---|---|---|---|
| Legal | 440.2 | 13.4 | -7.9 | -5.6 |
| Government & regulatory affairs | 13.8 | 286.7 | 157.8 | 0.4 |
| Trade executives | -23.6 | 93.5 | 211.6 | -4.8 |
| Pipeline employees | -4.8 | -5.9 | -6.5 | 172.4 |



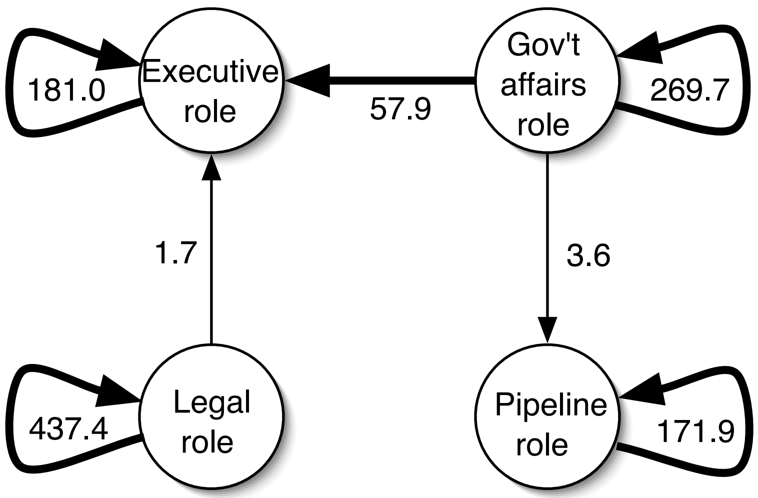

- Mostly communication within roles
- Some asymmetric exchanges
- Negative values hinder simple interpretation

# Communication Patterns




Nonnegative variant
NN-ASALSAN

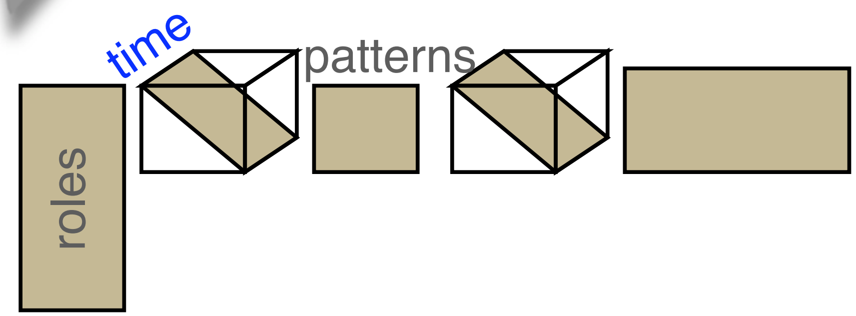
| | Legal | Gov't affairs | Trade execs | Pipeline |
|---|---|---|---|---|
| Legal | 437.4 | 0 | 1.7 | 0 |
| Government & regulatory affairs | 0 | 269.7 | 57.9 | 3.6 |
| Trade executives | 0 | 0 | 181.0 | 0 |
| Pipeline employees | 0 | 0 | 0 | 171.9 |

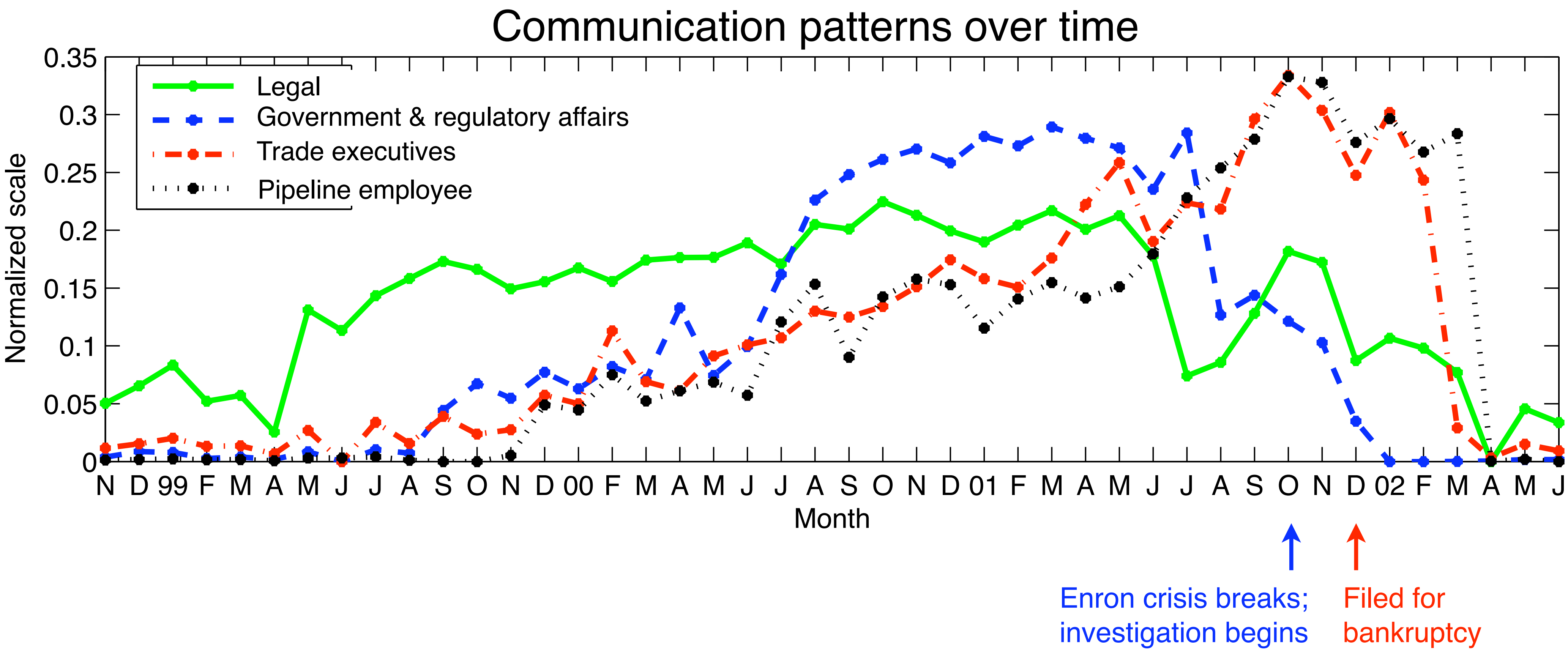



- Simplified graph
- Easier to understand

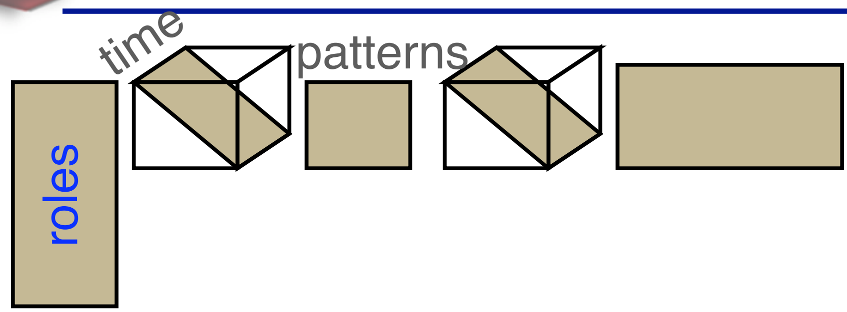
Sandia National Laboratories

# Temporal Patterns



# Precision of Categorization




| True label | Highest score | 1st and 2nd highest score |
|---|---|---|
| **ASALSAN** | | |
| Executive | 75% | 95% |
| Legal | 73% | 80% |
| Pipeline | 62% | 77% |
| Overall | 73% | 89% |
| **NN-ASALSAN** | | |
| Executive | 73% | 93% |
| Legal | 73% | 87% |
| Pipeline | 62% | 85% |
| Overall | 71% | 90% |

# Summary

- ASALSAN algorithm
  - New procedure for finding $\mathbf{A}$
  - Newton step for finding $\mathbf{D}$

- NN-ASALSAN algorithm
  - Nonnegative version based on multiplicative updates

- Modifications to handle large data arrays
  - Compression

- Novel approach to social network analysis using DEDICOM
  - Roles of employees
  - Communication patterns among roles and over time

- Future research
  - Constrained DEDICOM

# More Information

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

- MATLAB Tensor Toolbox:
  - http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox
  - Paper in ACM Trans. Math. Softw.
  - Paper to appear soon in SISC