SAND2007-4390C

An Overview of Tensor Decompositions and their Applications

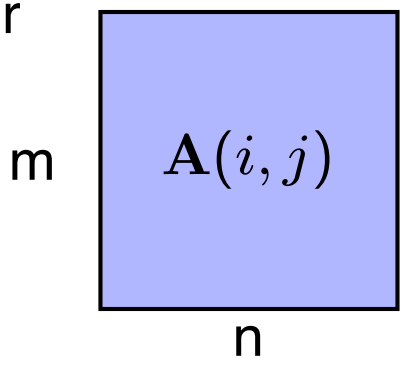
Brett W. Bader & Tamara G. Kolda
Sandia National Laboratories

ICIAM 2007
July 19, 2007

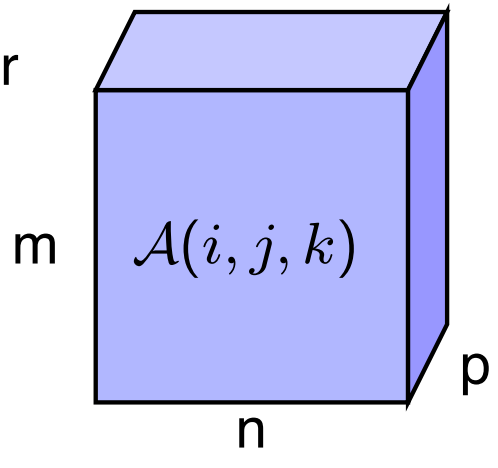
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Tensors

2nd order Tensor (Matrix)



3rd order Tensor

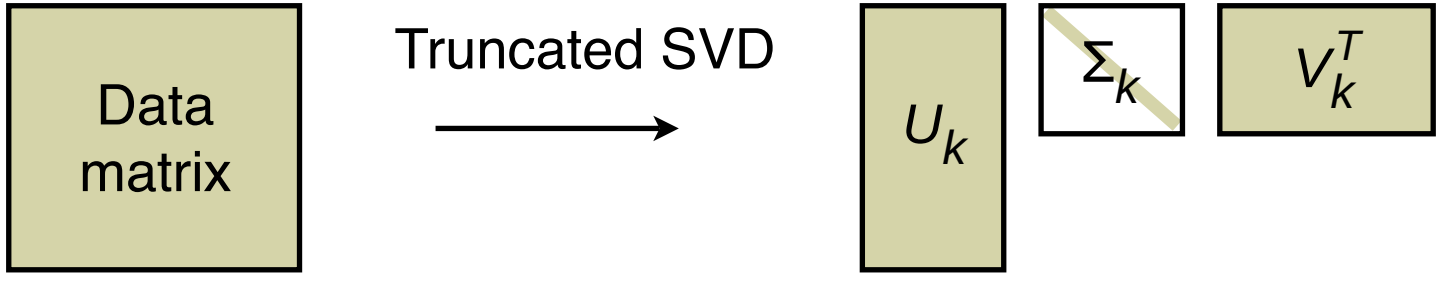


- Definition:
  - Multidimensional array
  - N-way array
  - Informally, data with more than 2 subscripts (but could refer to vectors and matrices)

- The order of a tensor is the number of dimensions (or “ways” or “modes”)
  - Scalar = tensor of order 0
  - Vector = tensor of order 1
  - Matrix = tensor of order 2

Matrix Decompositions

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



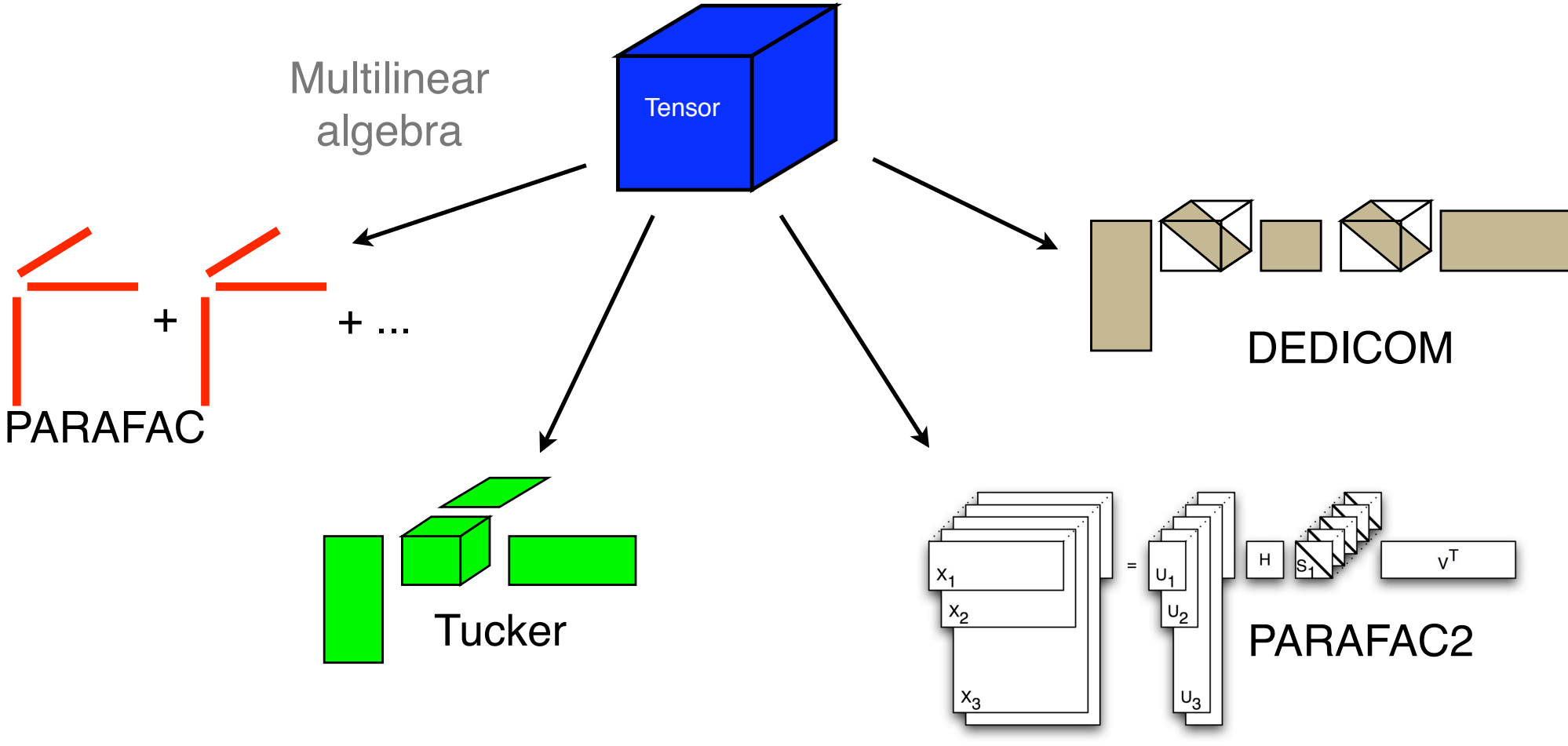
Best rank-*k* matrix filters out noise and captures "latent" information, which improves certain data mining tasks

Examples:

- Latent Semantic Analysis
- Text Analysis
- Web search
- Reduced Order Models (POD)

But we may be ignoring useful information in the data!

Tensor Decompositions



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

Mathematical Notation

- Scalars a
- Vectors $\mathbf{a}$
- Matrices $\mathbf{A}$
- Tensors (3-way array) $\mathcal{D}$ $\mathcal{X}$
- Special symbols
  - Outer (tensor) product

$$\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^T$$

  - Kronecker product

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

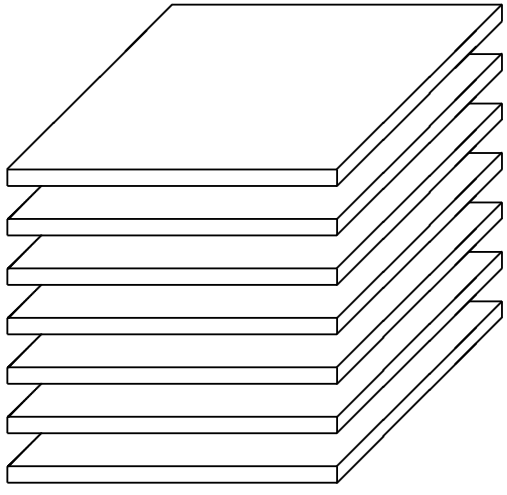
  - Khatri-Rao product (columnwise Kronecker)

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \dots & \mathbf{a}_n \otimes \mathbf{b}_n \end{bmatrix}$$

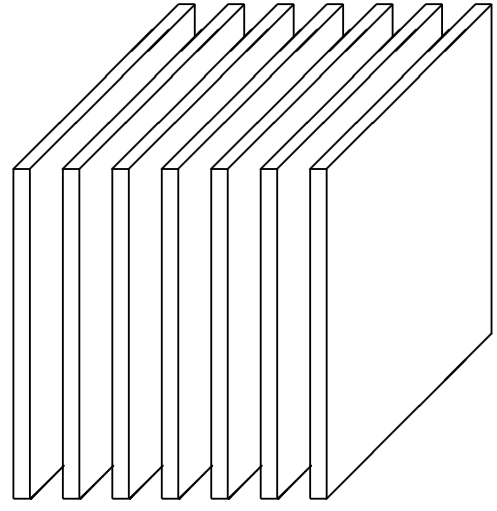
  - Hadamard product (elementwise)

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

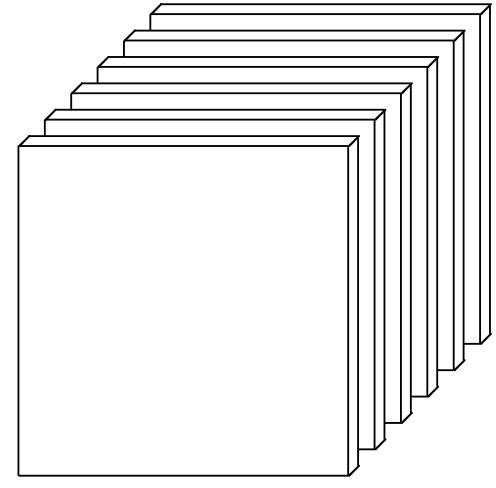
Slices and Fibers



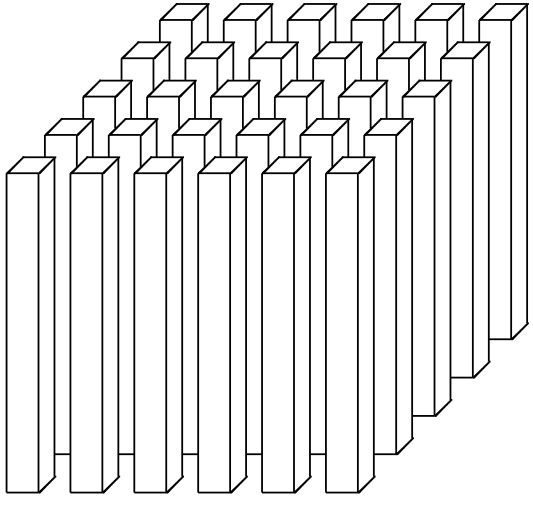
Horizontal $\mathcal{A}(i,:,:)$



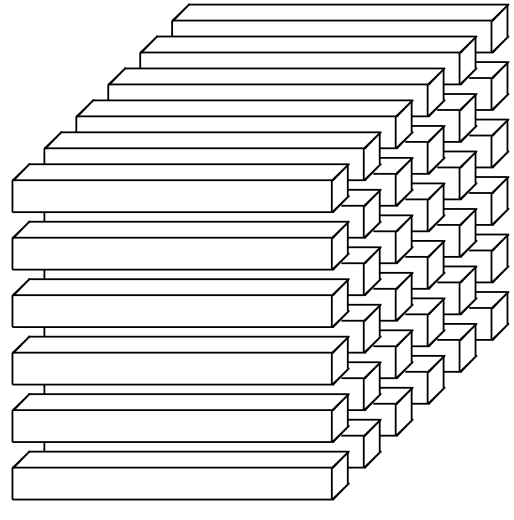
Lateral $\mathcal{A}(:,j,:)$



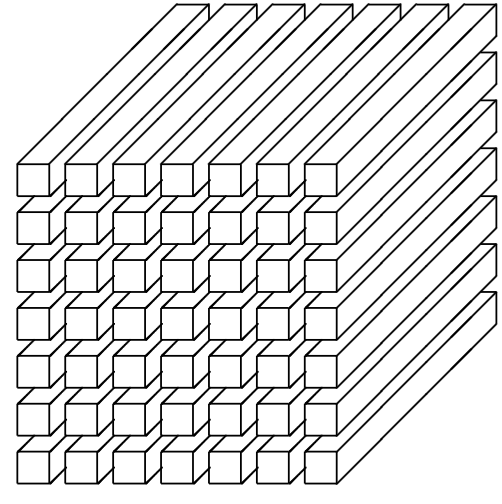
Frontal $\mathcal{A}(:,:,k)$



Mode-1 — Columns
$\mathcal{A}(:,j,k)$



Mode-2 — Rows
$\mathcal{A}(i,:,k)$

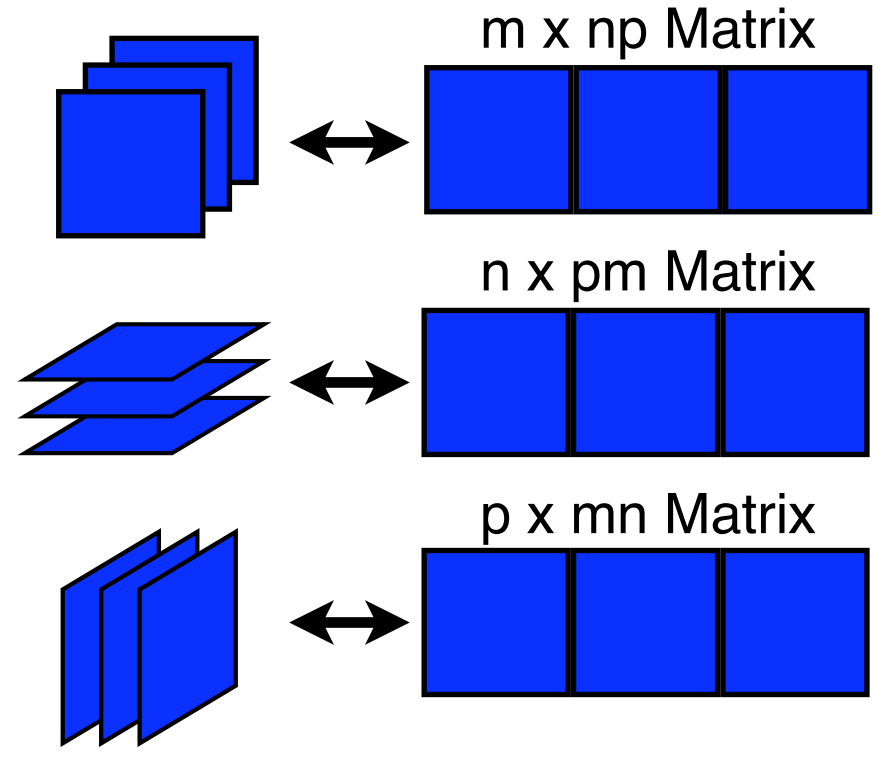


Mode-3 — Tubes
$\mathcal{A}(i,j,:)$

Sandia National Laboratories

Matricize: Convert a Tensor to a Matrix

- Convert a matrix to a tensor
- Also called “unfolding” or “flattening”
- $\mathbf{A}_{(n)}$= matrix form of a tensor where the nth dimension mapped to the row index of the matrix



- Many schemes are possible:
  - Any set of indices can be mapped to the rows
  - The remaining indices mapped to the columns
  - Order of the columns must be consistent
- Reverse matricize possible (the map is stored)

Multiplication with Tensors

- Many ways to multiply tensors!!

- Need to specify...
  - Tensor mode (i.e., dimension) to be multiplied
  - Whether result should be “contracted” (tensor-vector product)
  - How the dimensions of the result should be arranged

- Called n-mode product: specify the mode of the tensor involved in the multiplication, e.g., 1-mode product involves the tensor “fibers” in the 1st mode (i.e., columns)

$$\mathcal{C} = \mathcal{A} \times_n \mathbf{B}$$

- Commonly implemented with matricize and linear algebra

$$\mathbf{C}_{(n)} = \mathbf{B}\mathbf{A}_{(n)}$$

Typically, multiplication with tensors (and associated operations) is dominant cost of algorithm. Thus, we structure classes to efficiently handle these operations.

Tensor-Matrix Multiplication

- n-mode product with matrices does not reduce order of result but may change the dimension

- $\mathcal{B} = \mathcal{A} \times_1 \mathbf{U}$ (multiply A by U in the first mode)

$$\mathcal{B}(i,j,k) = \sum_{i'=1}^{m} \mathcal{A}(i',j,k) \cdot \mathbf{U}(i,i')$$

- $\mathcal{C} = \mathcal{A} \times_1 \mathbf{U} \times_2 \mathbf{V}$ (multiply A by U in the first mode, then by V in the second mode)

$$\mathcal{C}(i,j,k) = \sum_{i'=1}^{m} \sum_{j'=1}^{n} \mathcal{A}(i',j',k) \cdot \mathbf{U}(i,i') \cdot \mathbf{V}(j,j')$$

Matrix SVD in Tensor Notation

- Notation generalizes matrix-matrix products:

$$\mathbf{A} \times_1 \mathbf{U} = \mathbf{U}\mathbf{A}$$

$$\mathbf{A} \times_2 \mathbf{V} = \mathbf{A}\mathbf{V}^T$$

- Can express matrix SVD with n-mode products:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Sigma} \times_1 \mathbf{U} \times_2 \mathbf{V}.$$

Tensor-Vector Multiplication

- Order of result reduced ("squeeze" singletons)

bar indicates contracted product

- $\mathbf{B} = \mathcal{A} \bar{\times}_1 \mathbf{u}$ (multiply A by u in the first mode)

$$\mathbf{B}(j,k) = \sum_{i'=1}^{m} \mathcal{A}(i',j,k) \cdot \mathbf{u}(i')$$

- $\mathbf{c} = \mathcal{A} \bar{\times}_1 \mathbf{u} \bar{\times}_2 \mathbf{v}$ (multiply A by u in the first mode, then by v in the second mode)

$$\mathbf{c}(k) = \sum_{i'=1}^{m} \sum_{j'=1}^{n} \mathcal{A}(i',j',k) \cdot \mathbf{u}(i') \cdot \mathbf{v}(j')$$

Building Blocks: Rank-1 Tensors

- Matrix (2nd order tensor)

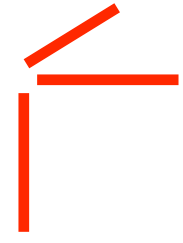
$$\mathbf{B} = u \circ v = uv^T$$

$$\mathbf{B}(i,j) = u(i) \cdot v(j)$$

- 3rd order tensor

$$\mathcal{C} = u \circ v \circ w$$

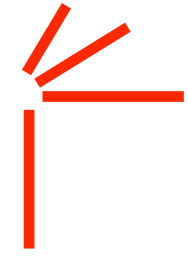
$$\mathcal{C}(i,j,k) = u(i) \cdot v(j) \cdot w(k)$$



- 4th order tensor

$$\mathcal{D} = u \circ v \circ w \circ x$$

$$\mathcal{D}(i,j,k,l) = u(i) \cdot v(j) \cdot w(k) \cdot x(l)$$



- Alternate notation:

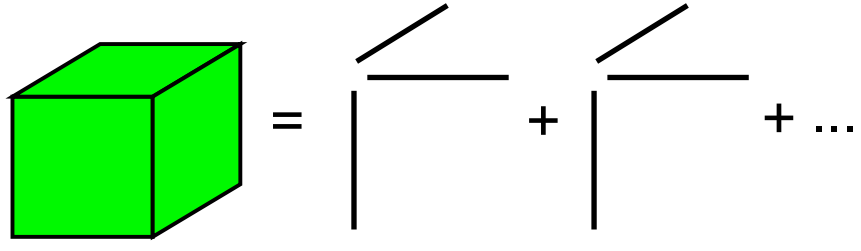
$u \circ v$ (usually outer product)

$u \otimes v$ (usually Kronecker product)

Tensor Decompositions

CANDECOMP-PARAFAC (CP) Decomposition

$$\mathcal{T} = \sum_{i=1}^{K} \lambda_i \mathbf{U}_i^{(1)} \circ \mathbf{U}_i^{(2)} \circ \mathbf{U}_i^{(3)}$$



- Invented by both Carroll and Chang (1970) and Harshman (1970)
- Sum of rank-1 tensors
- The columns of each U are not necessarily orthogonal
- If K is minimal, then K is the rank of the tensor
- The rank of a tensor can be greater than min(m,n,p)
- Unique decomposition that is not subject to rotation

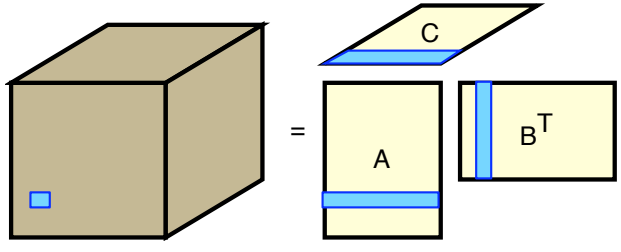
`T = cp_tensor(lambda,U)` creates a `cp_tensor` object. Here `lambda` is a K-vector and `U` is a cell array whose nth entry is the matrix $U^{(n)}$ with K columns.

PARAFAC

- Many ways to write the mathematical model:

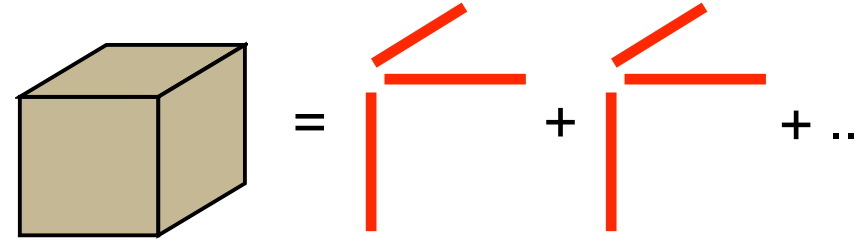
Scalar form

$$x_{ijk} \approx \sum_{i=1}^{r} a_{ir} b_{jr} c_{kr}$$



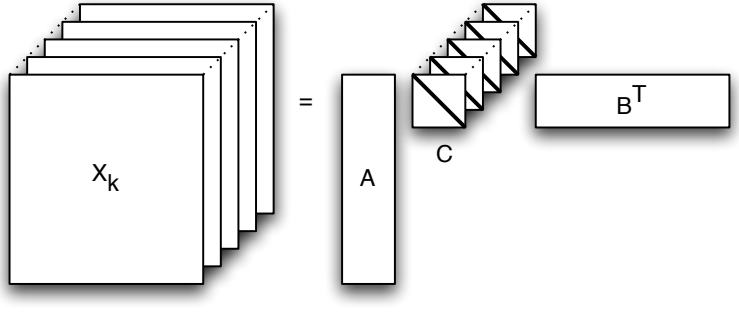
Outer product form

$$\mathcal{X} \approx \sum_{i=1}^{r} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$$



Tensor slice form

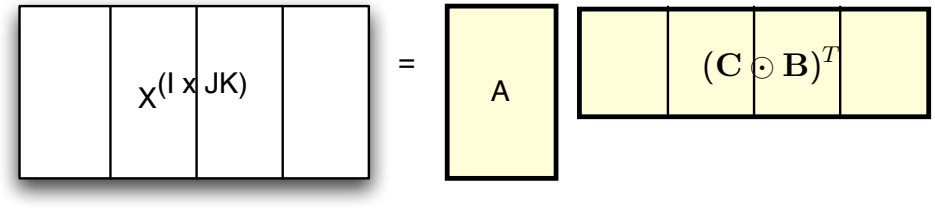
$$\mathbf{X}_k \approx \mathbf{A}\ \mathrm{diag}(\mathbf{c}_{k:})\mathbf{B}^T$$



Matrix form

$$\mathbf{X}^{I \times JK} \approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T$$

Matricized array

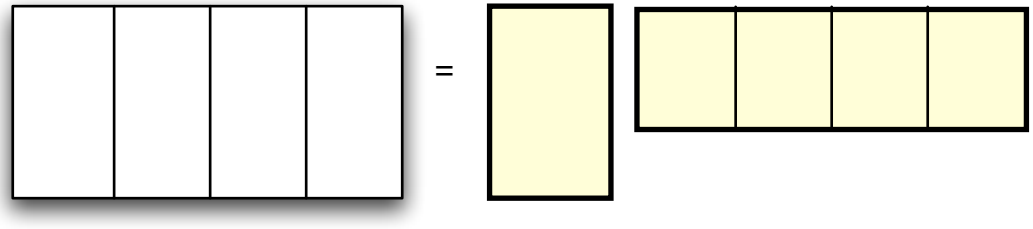


Sandia National Laboratories

PARAFAC Algorithm

Typically solved by Alternating Least Squares

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \mathbf{X}^{I\times JK} - \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T \right\|_F = \left\| \mathbf{X}^{J\times IK} - \mathbf{B}(\mathbf{C}\odot\mathbf{A})^T \right\|_F = \left\| \mathbf{X}^{K\times IJ} - \mathbf{C}(\mathbf{B}\odot\mathbf{A})^T \right\|_F$$



Minimize over $\mathbf{A}, \mathbf{B}, \mathbf{C}$ using least-squares solution:

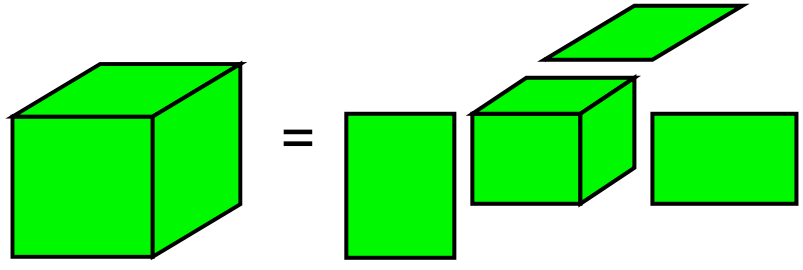
$$\mathbf{A} \leftarrow \mathbf{X}^{I\times JK}\mathbf{Z}^{\dagger}, \qquad \mathbf{Z} = (\mathbf{C}\odot\mathbf{B})^T$$

$$\mathbf{B} \leftarrow \mathbf{X}^{J\times IK}\mathbf{Z}^{\dagger}, \qquad \mathbf{Z} = (\mathbf{C}\odot\mathbf{A})^T$$

$$\mathbf{C} \leftarrow \mathbf{X}^{K\times IJ}\mathbf{Z}^{\dagger}, \qquad \mathbf{Z} = (\mathbf{B}\odot\mathbf{A})^T$$

Tucker Decomposition

$$\mathcal{T} = \lambda \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(1)} \times_3 \mathbf{U}^{(3)}$$

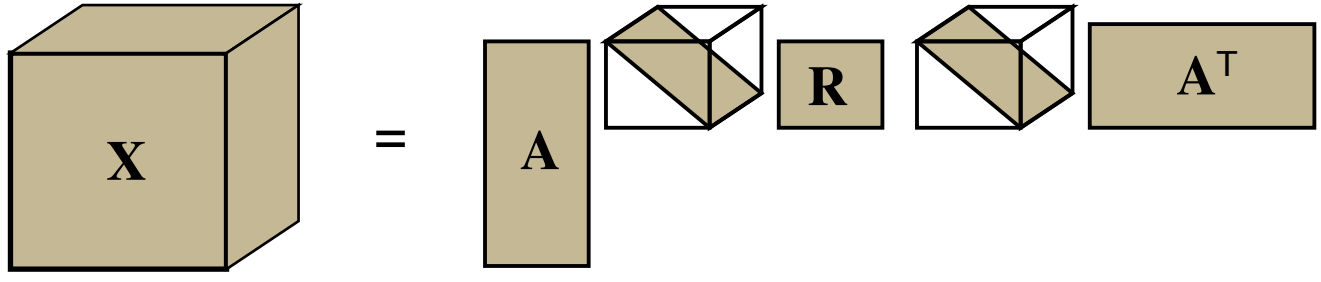


- Invented by Tucker (1966)
- Sum of rank-1 tensors
- Columns of each $U^{(i)}$ are orthogonal
- Core tensor is dense
- Associated with n-mode ranks, i.e., each dimension has its own rank
- The HOSVD computes a Tucker decomposition

`T = tucker_tensor(lambda,U)` where `lambda` is a $K_1 \times K_2 \times \cdots \times K_N$ tensor and `U` is a cell array whose nth entry is a matrix with K_n columns.

Three-way DEDICOM

$$\mathbf{X}_x = \mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\mathbf{A}^\mathsf{T} \qquad k = 1, \ldots, \mathcal{K}$$



- Introduced by Harshman (1978)
- Sum of rank-1 tensors
- Columns of A are not necessarily orthogonal
- Central matrix R contains asymmetric information from X
- Alternating algorithms, least-squares and approximations
- Early applications:
  - World trade (import/export matrices)
  - Car switching
- Variations: constrainted DEDICOM

Software packages

- Many packages available for various applications
  - MATLAB Tensor Toolbox, Version 2.2, Brett W. Bader and Tamara G. Kolda, Sandia National Laboratories, `http://csmr.ca.sandia.gov/%7Etgkolda/TensorToolbox/`
  - N-Way Toolbox for MATLAB [**?**], Version 3, Claus A. Andersson and Rasmus Bro, University of Copenhagen, Denmark, `http://www.models.kvl.dk/source/nwattoolbox/`.
  - MATLAB, Version 2006b, The Mathworks, Inc., `http://www.mathworks.com/`
  - Boost.Multiarray, The Boost Multidimensional Array Library, Version 1.31.0, Ronald Garcia, Jeremy Sick, and Andrew Lumsdaine, Indiana University, `http://www.boost.org/libs/multi+array/doc/index.html`
  - HTL, the HUJI Tensor Library, Version 0.04, Ron Zass, Hebrew University of Jerusalem, Israel, `http://www.cs.huji.ac.il/~zass/htl/`
  - FTensor, Version 1.1pre-25, Walter Landry, University of Utah and Uiversity of California San Diego, `http://www/oonumerics.org/FTensor/`
  - CuBatch [**?**], S. Gourvénec et al., `http://www.models.life.ku.dk/source/CuBatch/`
  - Multilinear Engine, Penti Paatero, University of Helsinki

Survey of Data Mining Applications

- Discussion tracking in email communications
- Social network analysis
- Cross-language information retrieval
- Web link analysis

Discussion Tracking in Enron Email using PARAFAC

*Brett W. Bader**, Michael W. Berry**, and Murray Browne**
*Sandia National Laboratories
**University of Tennessee

Text Mining Workshop 2007
SIAM Conference on Data Mining
April 28, 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.

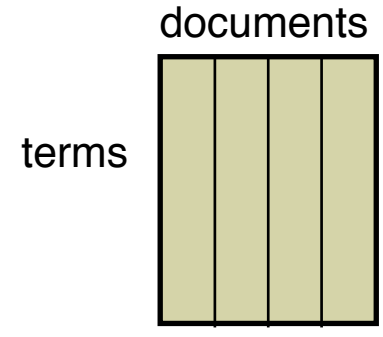
NNSA National Nuclear Security Administration

Sandia National Laboratories

Common Text Analysis Approach: Latent Semantic Indexing

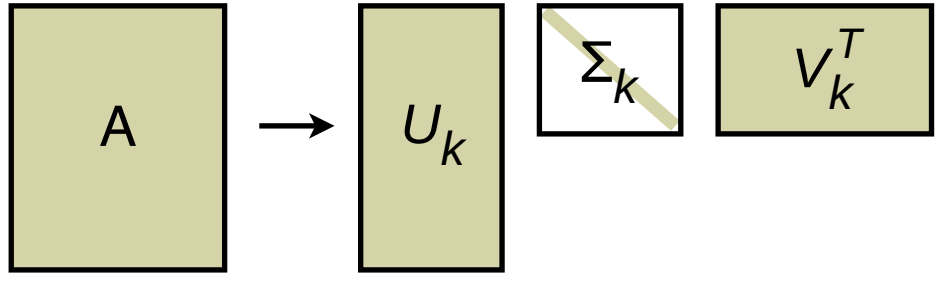
(Deerwester, Dumais, Furnas, Landauer, Harshman, 1990)

Replace term-document matrix with a lower rank matrix that captures "latent" information



Use truncated SVD to compute best rank-*k* matrix

$$A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \longrightarrow A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$



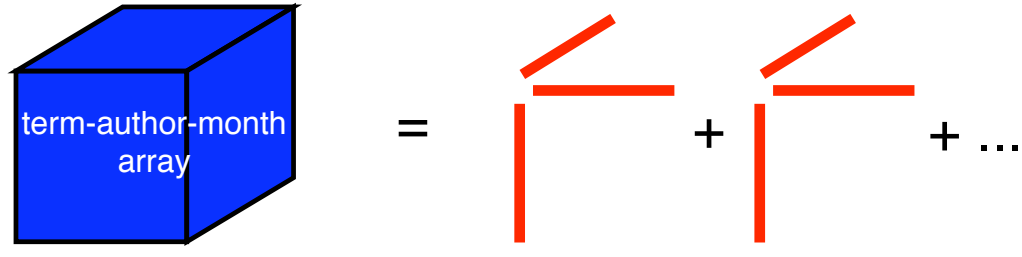
Dimension reduction filters out noise and captures latent information, which improves certain text mining tasks

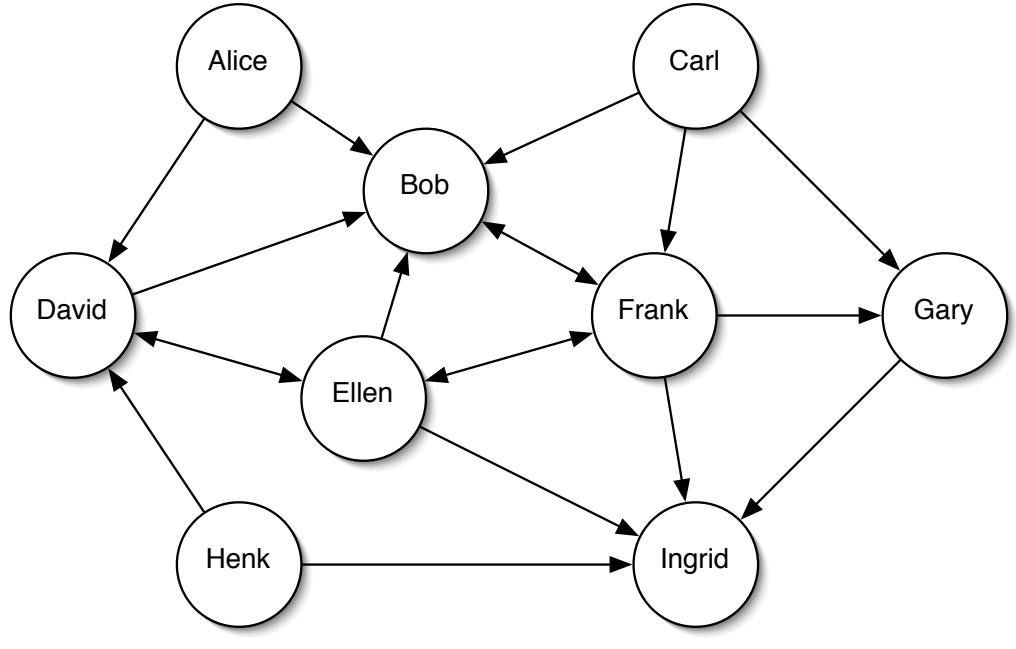
Recent interest in non-negative (parts-based) decompositions

Sandia National Laboratories

Objective

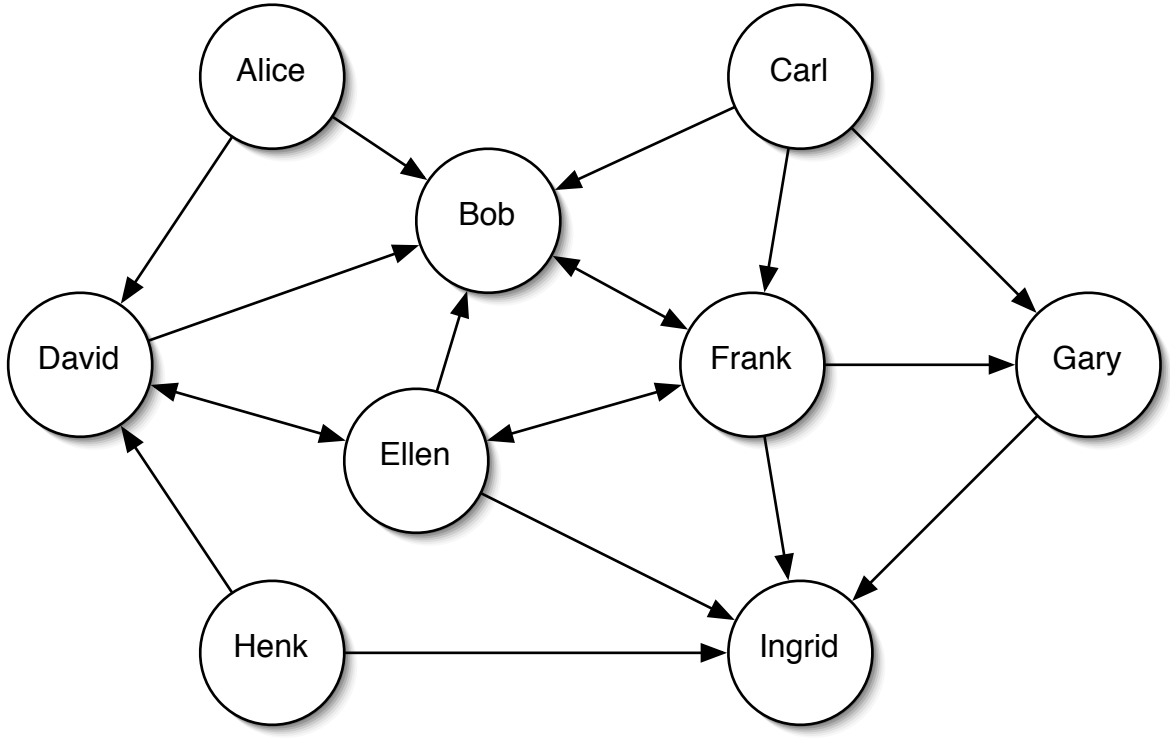
Use PARAFAC to extend LSI to analyze content of email communications over time





Sandia National Laboratories

Application: Email Surveillance

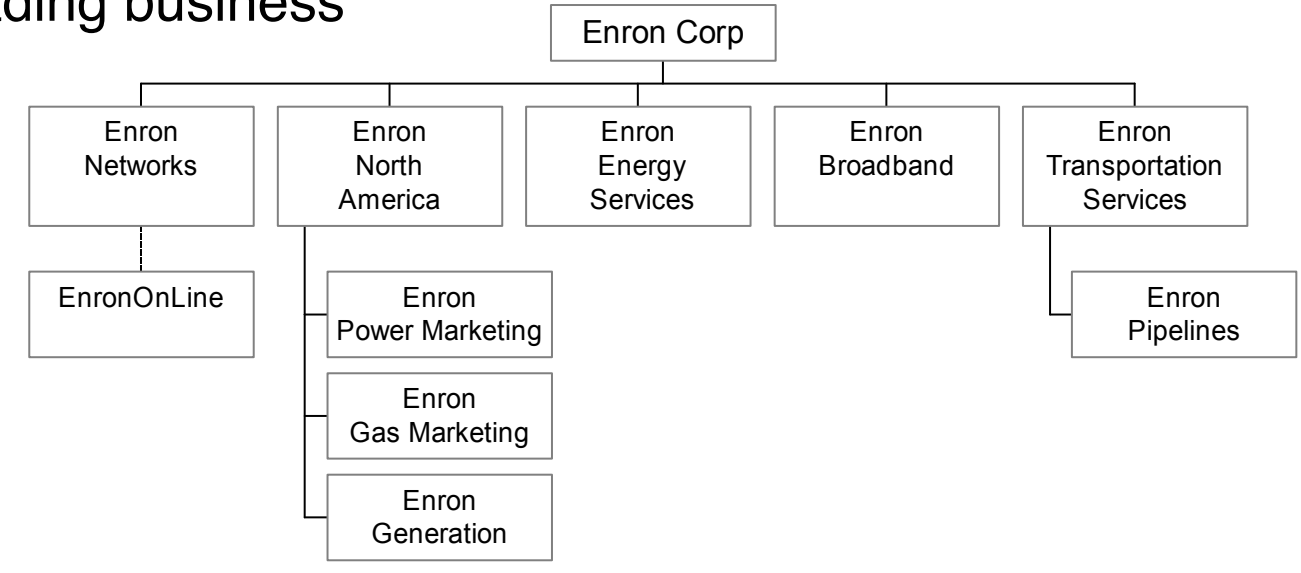


- Links are email communications

- What can we learn about their email conversations?
  - What are the major topics of conversations?
  - Who are the major participants?
  - When are they taking place?

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

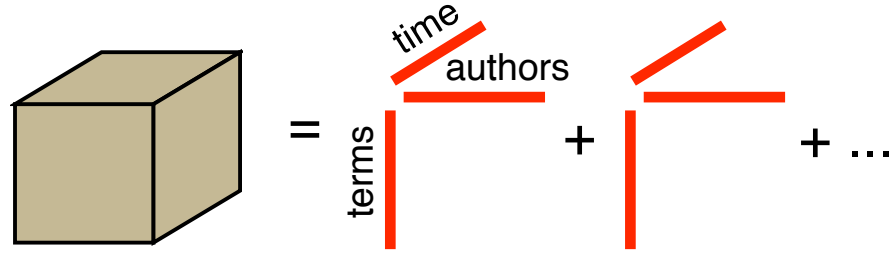
- FERC collected email of ~150 employees as evidence
  - Included emails saved in inbox, sent items, deleted items, and all other folders

- 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

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

Text Analysis Experiment

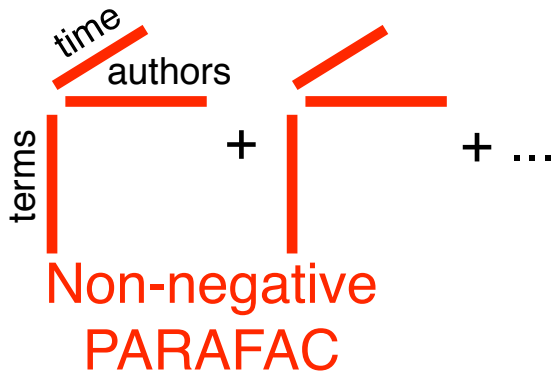
- Computed rank-25 models:
  - PARAFAC
  - Non-negative PARAFAC



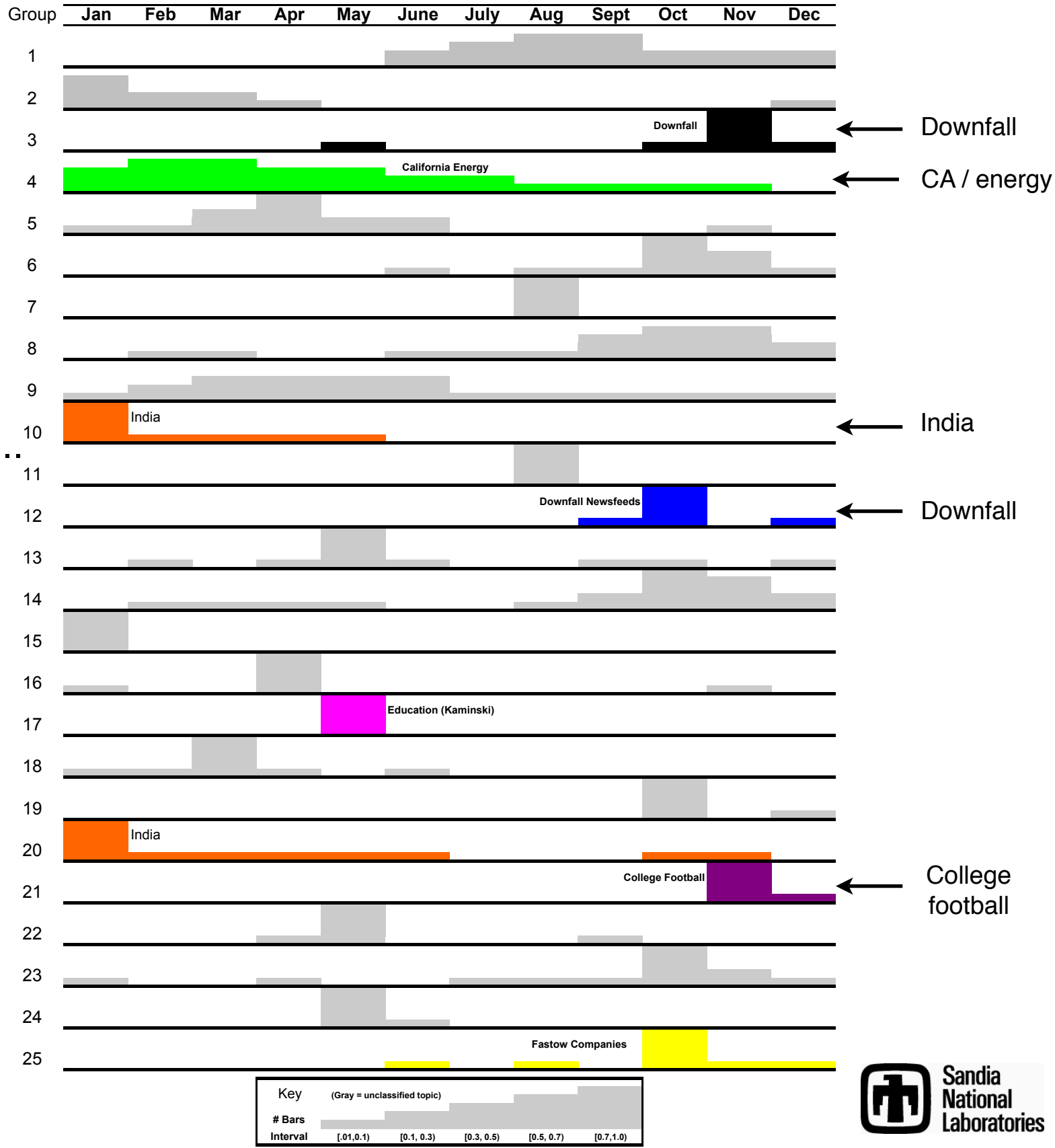
25 groups $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$
corresponding to 25 discussions

- Relative residual error = 0.8904 and 0.8931, respectively

Topics over time



Non-negative PARAFAC



Downfall

CA / energy

India

Downfall

College football

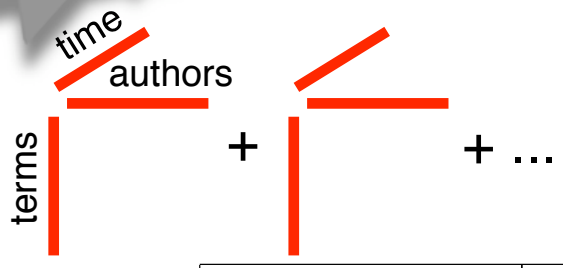
Kaminski / Education

India

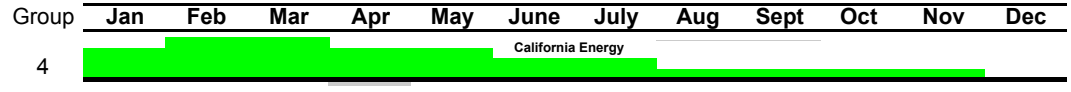
Fastow companies

Sandia National Laboratories

Conversation Topics of Employees

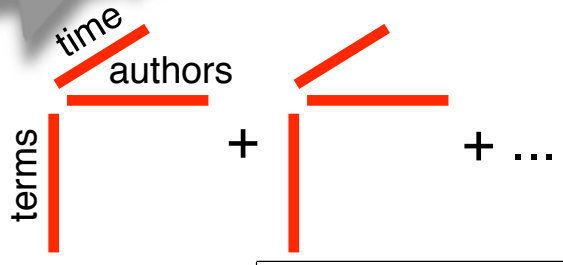


California Energy

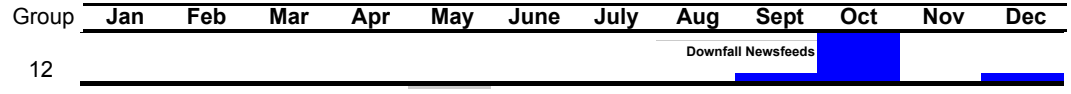


Terms		Authors	
Score	Noun	Score	Employee
Factor 4			
0.094	california	0.497	James Steffes (james.steffes) VP Government Affairs
0.087	dasovich	0.430	Steven Kean (steven.kean) VP Chief of Staff
0.079	jeff	0.413	Jeff Dasovich (jeff.dasovich) Employee Government Relationship Executive
0.077	shapiro	0.319	Richard Sanders (richard.sanders) VP Enron Wholesale Services
0.076	steffes	0.219	Richard Shapiro (richard.shapiro) VP Regulatory Affairs
0.075	richard	0.194	Elizabeth Sager (elizabeth.sager) VP and Asst Legal Counsel ENA Legal
0.073	kean	0.187	Mark Haedicke (mark.haedicke) Managing Director ENA Legal
0.072	edison	0.171	Drew Fossum (drew.fossum) VP Transwestern Pipeline Company (ETS)?
0.067	utilities	0.152	Philip Allen (phillip.allen) VP West Desk Gas Trading
0.066	power	0.134	Kay Mann (kay.mann) Lawyer
0.065	sanders	0.125	Mark Taylor (mark.taylor) Manager Financial Trading Group ENA Legal
0.064	mara	0.100	John Arnold (john.arnold) VP Financial Enron Online
0.063	james	0.097	Margaret Carson (margaret.carson) Employee Corporate and Environmental Policy*
0.062	development	0.095	Kevin Presto (kevin.presto) VP East Power Trading
0.061	governor	0.085	Vince Kaminski (vince.kaminski) Manager Risk Management Head
0.061	vicki	0.081	David Delainey (david.delainey) CEO ENA and Enron Energy Services
0.058	energy	0.072	Rick Buy (rick.buy) Manager Chief Risk Management Officer
0.057	kaufman	0.069	Sara Shackleton (sara.shackleton) Employee ENA Legal
0.055	mccubbin	0.060	Kate Symes (kate.symes) Employee
0.055	kingerski	0.059	Gerald Nemec (gerald.nemec) N/A
0.055	utility	0.055	Larry Campbell (larry.campbell) Employee Senior Specialist
0.055	sharp	0.055	Michael Grigsby (mike.grigsby) Director West Desk Gas Trading
0.055	market	0.054	Dan Hyvl (dan.hyvl) Employee
0.054	electricity	0.054	Mike McConnell (mike.mcconnell) Executive VP* Global Markets
0.054	alan	0.050	Bruce Lundstrom (bruce.lundstrom) N/A

Conversation Topics of Employees

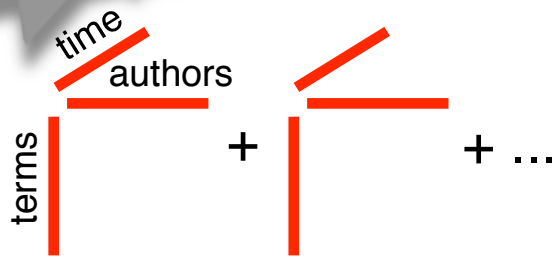


Enron crisis and downfall newsfeeds

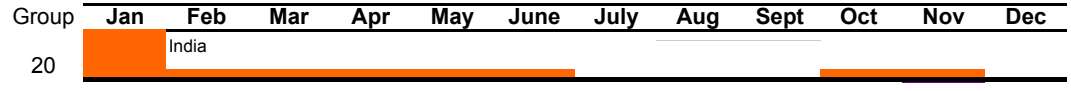


Terms		Authors	
Score	Noun	Score	Employee
Factor 12			
0.072	amat	0.790	Teb Lokey (teb.lokey) Manager Regulatory Affairs
0.063	skilling	0.480	Shelley Corman (shelley.corman) VP Regulatory Affairs
0.062	billion	0.256	Darrell Schoolcraft (darrell.schoolcraft) Employee Gas Control (ETS)
0.055	hng	0.176	Tana Jones (tana.jones) Employee Financial Trading Group ENA Legal
0.054	jeff	0.165	Louise Kitchen (louise.kitchen) President Enron Online
0.051	profits	0.076	Daron Giron (c..giron) Employee
0.051	percent	0.073	Phillip Love (m..love) N/A
0.050	california	0.047	Mark Whitt (mark.whitt) Director Marketing
0.050	electricity	0.031	Michael Grigsby (mike.grigsby) Director West Desk Gas Trading
0.049	blair	0.027	James Derrick (james.derrick) In House Lawyer
0.049	gest	0.026	Jay Reitmeyer (jay.reitmeyer) Associate Eastern Rockies Natural Gas Trader
0.047	teb	0.024	Lynn Blair (lynn.blair) Employee Northern Natural Gas Pipeline (ETS)
0.046	wall	0.024	Benjamin Rogers (benjamin.rogers) Employee Associate
0.043	lokey	0.021	Bruce Lundstrom (bruce.lundstrom) N/A
0.043	energy	0.021	Steven Kean (j..kean) VP Chief of Staff
0.043	customers	0.020	Stacey White (w..white) N/A
0.042	power	0.020	Jeff Dasovich (jeff.dasovich) Employee Government Relationship Executive
0.042	lay	0.019	James Steffes (d..steffes) VP Government Affairs
0.042	deregulation	0.018	John Arnold (john.arnold) VP Financial Enron Online
0.041	virgilio	0.018	Joe Quenet (joe.quenet) Trader
0.041	coale	0.018	Jeffrey Shankman (a..shankman) President Enron Global Markets
0.041	street	0.017	xxx Harris (j.harris) xxx
0.041	plants	0.013	Kim Ward (kim.ward) Manager West Gas Origination
0.041	million	0.011	Kenneth Lay (kenneth.lay) CEO
0.041	stock	0.010	Bill Williams (bill.williams) xxx

Conversation Topics of Employees



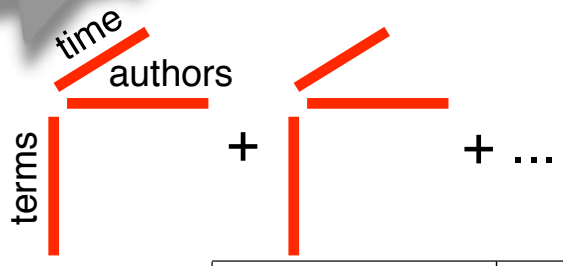
Dabhol Power Company (DPC) in India



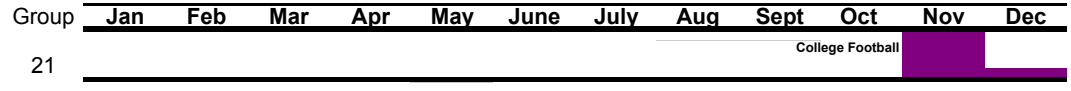
Terms		Authors	
Score	Noun	Score	Employee
Factor 20			
0.194	mseb	0.643	James Hughes (james.hughes) N/A
0.192	maharashtra	0.613	John Ambler (john.ambler) N/A
0.169	dabhol	0.384	Clay Harris (clay.harris) N/A
0.169	india	0.138	Bruce Lundstrom (bruce.lundstrom) N/A
0.160	hughes	0.118	Jeffrey Shankman (jeffrey.shankman) President Enron Global Markets
0.146	electricity	0.101	Sandeep Kohli (sandeep.kohli) N/A
0.142	power	0.055	Wade Cline (wade.cline) N/A
0.141	invoke	0.053	Mark Haedicke (mark.haedicke) Managing Director ENA Legal
0.141	dues	0.046	Steven South (steven.south) Director West Desk Gas Trading
0.135	dpc	0.046	Jeffery Skilling (jeff.skilling) CEO
0.126	billion	0.040	Hunter Shively (hunter.shively) VP
0.117	kean	0.038	Rob Gay (rob.gay) xxx
0.116	suppliers	0.037	Kevin Hyatt (kevin.hyatt) Director Asset Development TW Pipeline Business (ETS)
0.115	government	0.031	Philip Allen (phillip.allen) VP West Desk Gas Trading
0.114	defuse	0.030	Vince Kaminski (vince.kaminski) Manager Risk Management Head
0.112	rupees	0.027	John Lavorato (john.lavorato) CEO Enron America
0.105	decade	0.022	Mike McConnell (mike.mcconnell) Executive VP* Global Markets
0.104	unnamed	0.020	Kevin Ruscitti (kevin.ruscitti) Trader Central Desk Gas Trading
0.104	reuters	0.020	Charles Weldon (v.weldon) N/A
0.102	ambler	0.013	David Delainey (david.delainey) CEO ENA and Enron Energy Services
0.100	lenders	0.011	Debra Perlingiere (debra.perlingiere) Legal Specialist ENA Legal
0.096	liberalisation	0.010	Michelle Lokay (michelle.lokay) Admin. Asst. Transwestern Pipeline Company (ETS)
0.093	contractural		
0.087	adgas		
0.086	buys		

Sandia National Laboratories

Conversation Topics of Employees



College Football



Terms		Authors	
Score	Noun	Score	Employee
Factor 21			
0.189	bcs	0.737	Matthew Motley (matt.motley) Director
0.155	byu	0.606	Randall Gay (l..gay) West Desk Gas Trading
0.120	sooners	0.143	Craig Dean (craig.dean) Trader
0.119	frommelt	0.119	Mark Taylor (e.taylor) Manager Financial Trading Group ENA Legal
0.117	nebraska	0.091	Clint Dean (clint.dean) xxx
0.109	bowl	0.057	Kam Keiser (kam.keiser) Employee Gas
0.104	pooky	0.054	Eric Bass (eric.bass) Trader Texas Desk Gas Trading
0.102	gay	0.049	Thomas Martin (a..martin) VP
0.099	oklahoma	0.048	Cooper Richey (cooper.richey) Manager
0.097	big	0.045	Don Baughman (don.baughman) Trader
0.095	cougars	0.044	John Griffith (john.griffith) xxx
0.091	kathleen	0.044	Daren Farmer (j..farmer) Manager Logistics Manager
0.090	horns	0.044	Jim Schwieger (jim.schwieger) Trader Texas Desk Gas Trading
0.088	rooting	0.042	Kevin Hyatt (kevin.hyatt) Director Asset Development TW Pipeline Business (ETS)
0.086	fiesta	0.042	Albert Meyers (albert.meyers) Employee Specialist
0.085	tennessee	0.040	Bill Rapp (bill.rapp) N/A
0.085	texas	0.040	Michael Maggi (mike.maggi) Director
0.083	grigsby	0.039	Stanley Horton (stanley.horton) President Enron Gas Pipeline
0.081	longhorn	0.036	Cara Semperger (cara.semperger) Employee Senior Analyst Cash
0.081	oregon	0.033	Jeff King (jeff.king) Manager
0.080	longhorns	0.032	Sandra Brawner (f..brawner) Director
0.077	espn	0.031	Tom Donohoe (tom.donohoe) Trader Central Desk Gas Trading
0.077	miami	0.029	Jay Reitmeyer (jay.reitmeyer) Associate Eastern Rockies Natural Gas Trader
0.077	stanford	0.027	Jane Tholt (m..tholt) VP West Desk Gas Trading
0.077	large	0.027	Matthew Lenhart (matthew.lenhart) Analyst West Desk Gas Trading

Analysis of Latent Relationships in Semantic Graphs using DEDICOM

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

Workshop for Algorithms on Modern Massive Data Sets
June 24, 2006

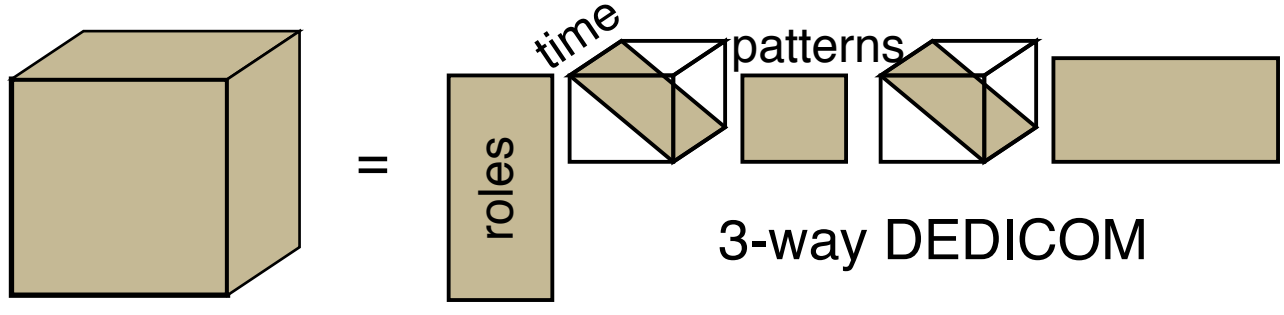
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

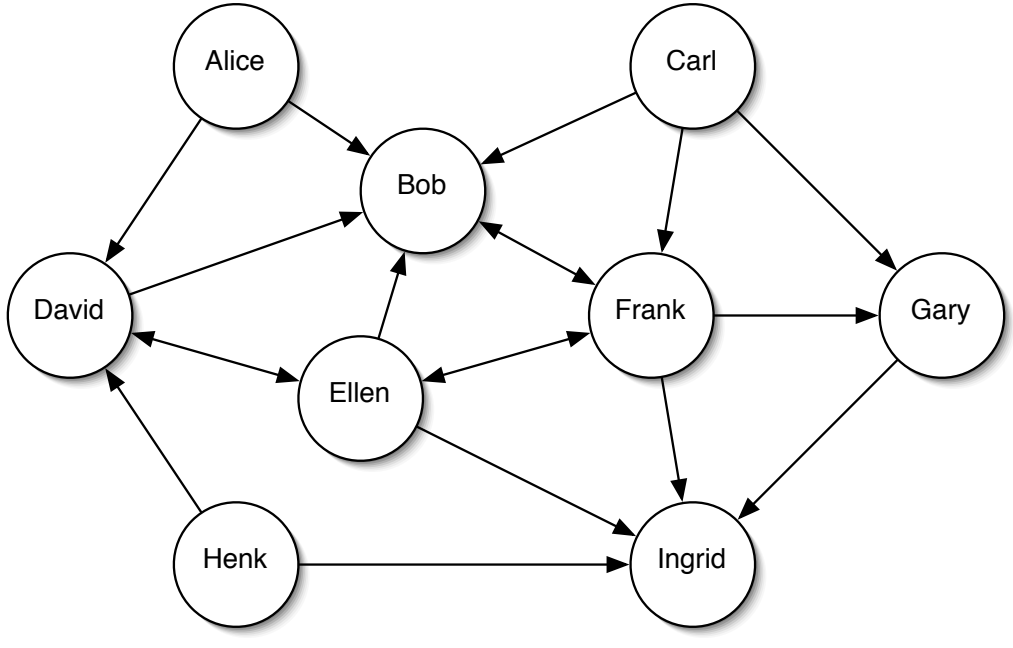
National Nuclear Security Administration

Sandia National Laboratories

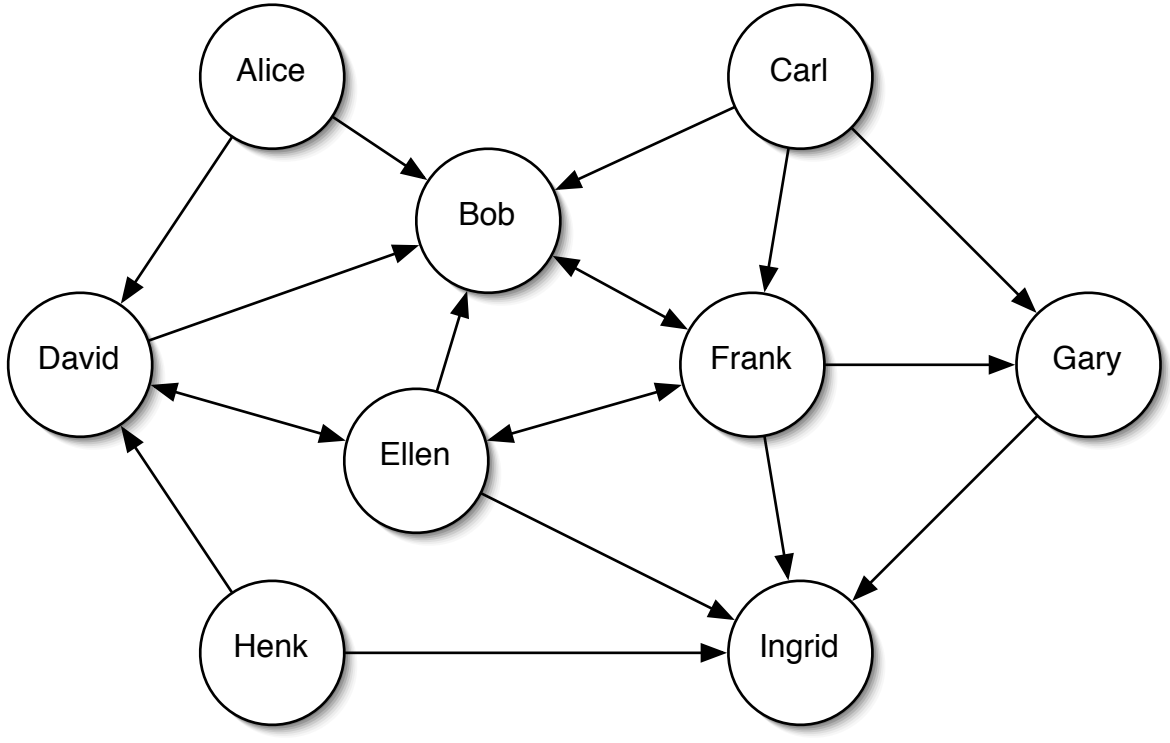
Objective

Use DEDICOM to analyze a semantic graph of email communications changing over time





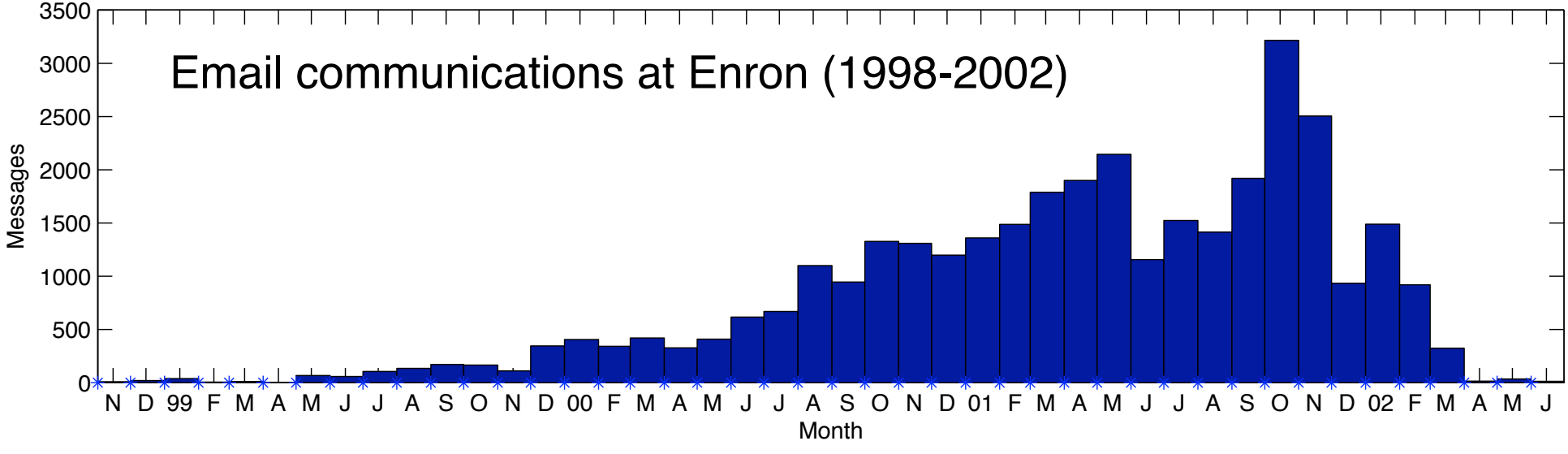
Application: Enron Email Analysis



- Links consist of email communications
- What can we learn about this network strictly from their communication patterns? (Social network analysis)

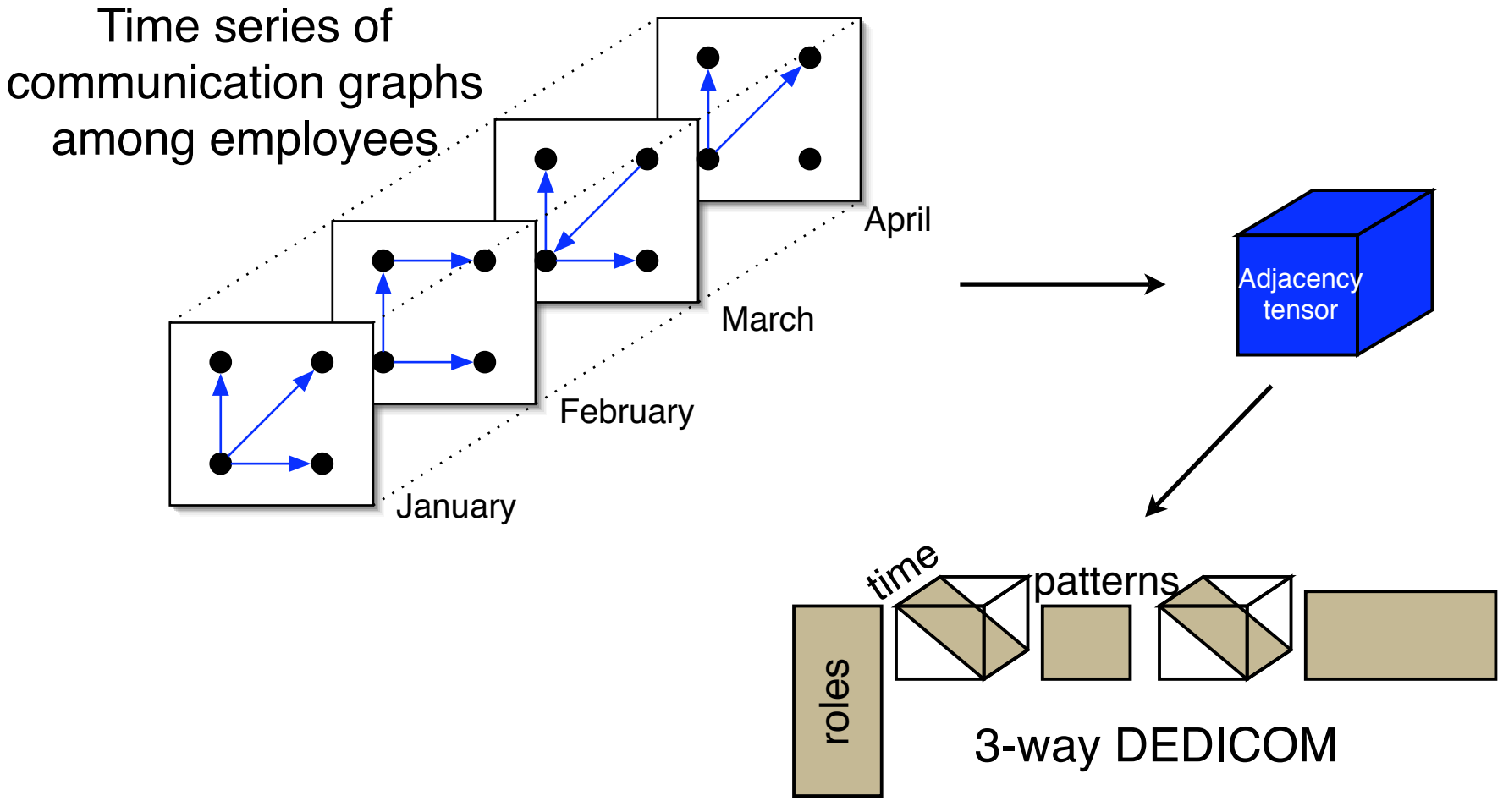
Smaller Enron Data Set

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



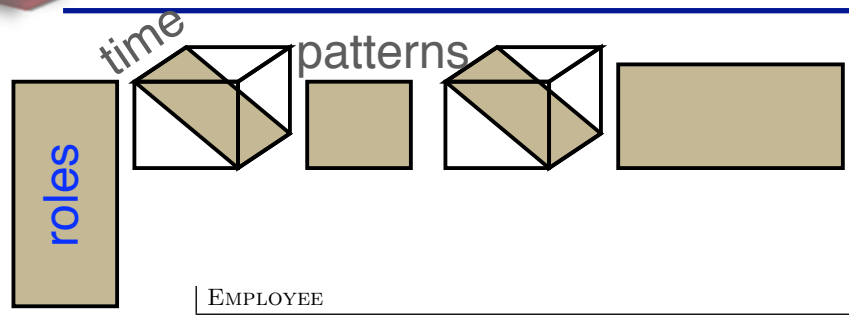
- Limited information on the 184 employees
- No org chart

Temporal Social Network Analysis



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

Roles of Employees

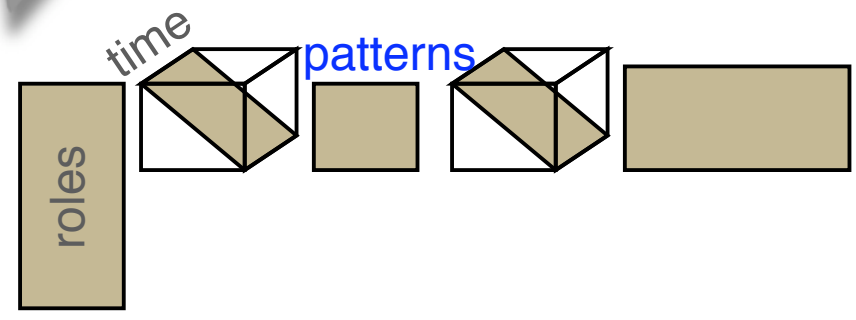


	Employee	1 (Legal)	2 (Gov't affairs)	3 (Trade execs)	4 (Pipeline)
Legal	T. Jones - Employee, Financial Trading Group (ENA Legal)	**0.64**	-0.01	0.02	-0.00
	S. Shackleton - Employee, ENA Legal	**0.45**	-0.00	-0.01	-0.00
	M. Taylor - Manager, Financial Trading Group ENA Legal	**0.37**	0.01	0.02	-0.00
	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**

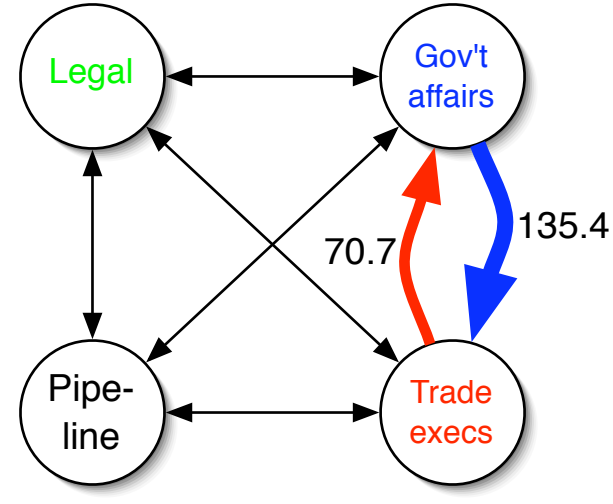
Identify shared characteristics to label group

Sandia National Laboratories

Communication Patterns

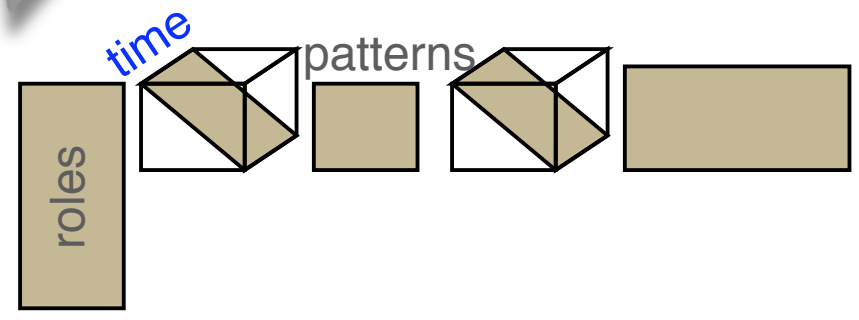


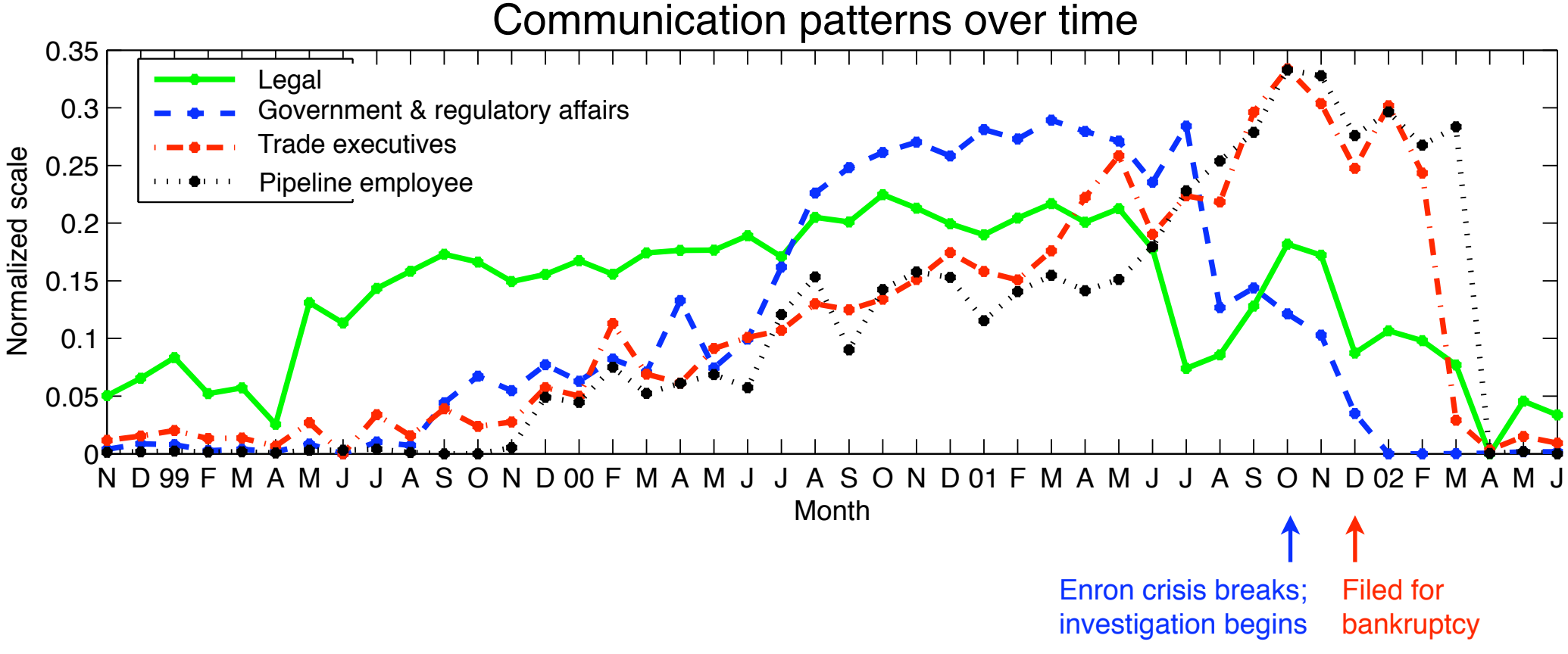
	Legal	Gov't affairs	Trade execs	Pipeline
Legal	440.2	1.6	-15.0	0.4
Government & regulatory affairs	1.6	278.3	135.4	1.6
Trade executives	-29.3	70.7	201.6	-6.2
Pipeline employees	1.4	-4.6	-7.5	172.3



- Mostly communication within roles
- Some large exchanges
- Negative values complicates interpretation
  - Non-negative factorization being investigated

Temporal Patterns





Sandia National Laboratories

Cross-language Information Retrieval using PARAFAC2

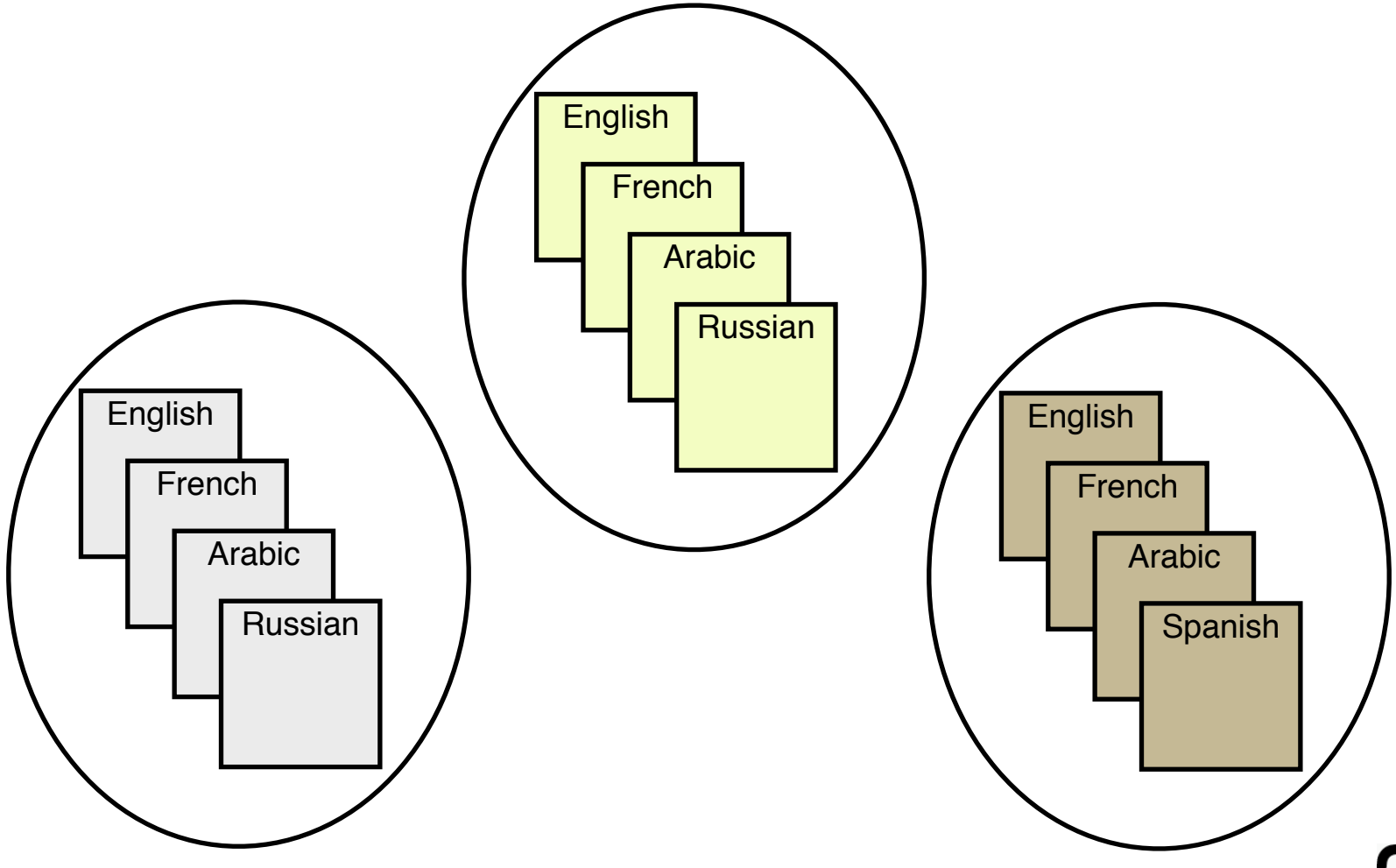
Peter Chew, Brett Bader, Tamara Kolda
Sandia National Laboratories

Knowledge Discovery and Data Mining (KDD)
August, 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.

Objective

Cluster documents that are
in different languages by topic



Bible as Parallel Corpus

The Bible has been translated into thousands of languages

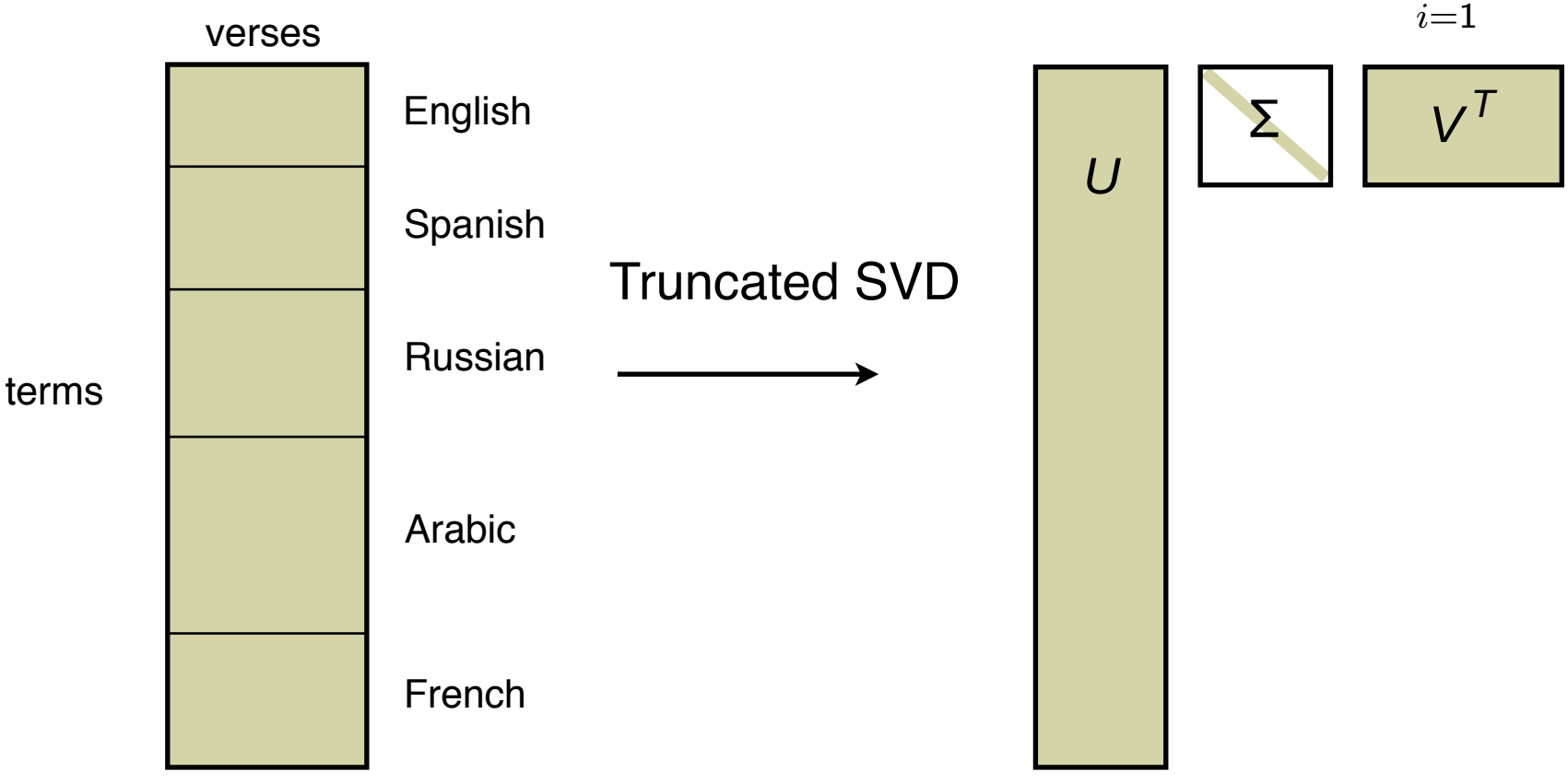
- Complete translations in 426 languages
- Partial translations in 2403 languages

Translation	Terms	Total Words
English (King James)	12,335	789,744
Spanish (Reina Valera 1909)	28,456	704,004
Russian (Synodal 1876)	47,226	560,524
Arabic (Smith Van Dyke)	55,300	440,435
French (Darby)	20,428	812,947

Latent Semantic Indexing

Term-by-verse matrix
for all language

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



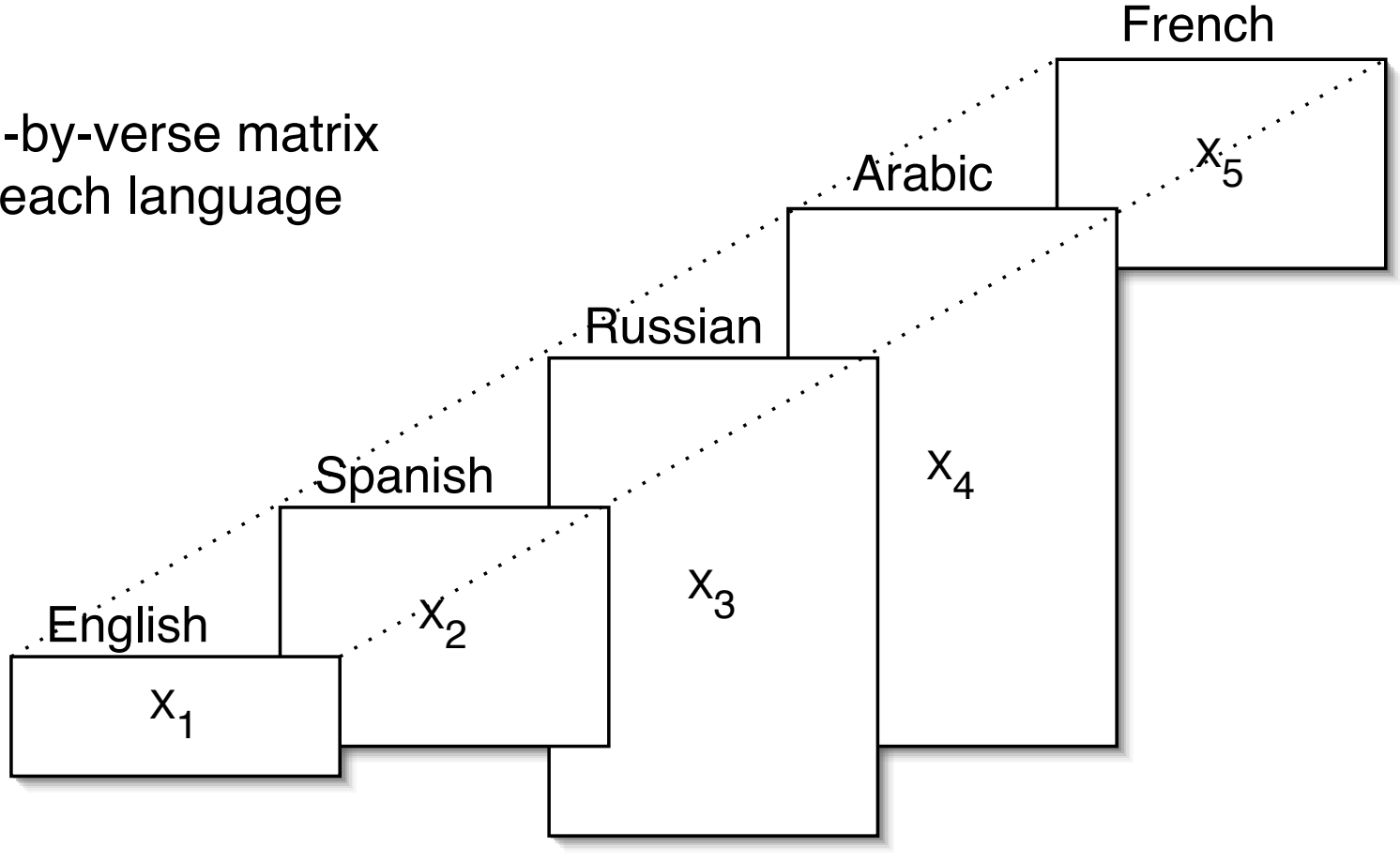
Project documents of interest into subspace of U
and compute cosine similarities

But documents tend to cluster by language, not by topic

Sandia National Laboratories

New Approach: Multi-matrix Array

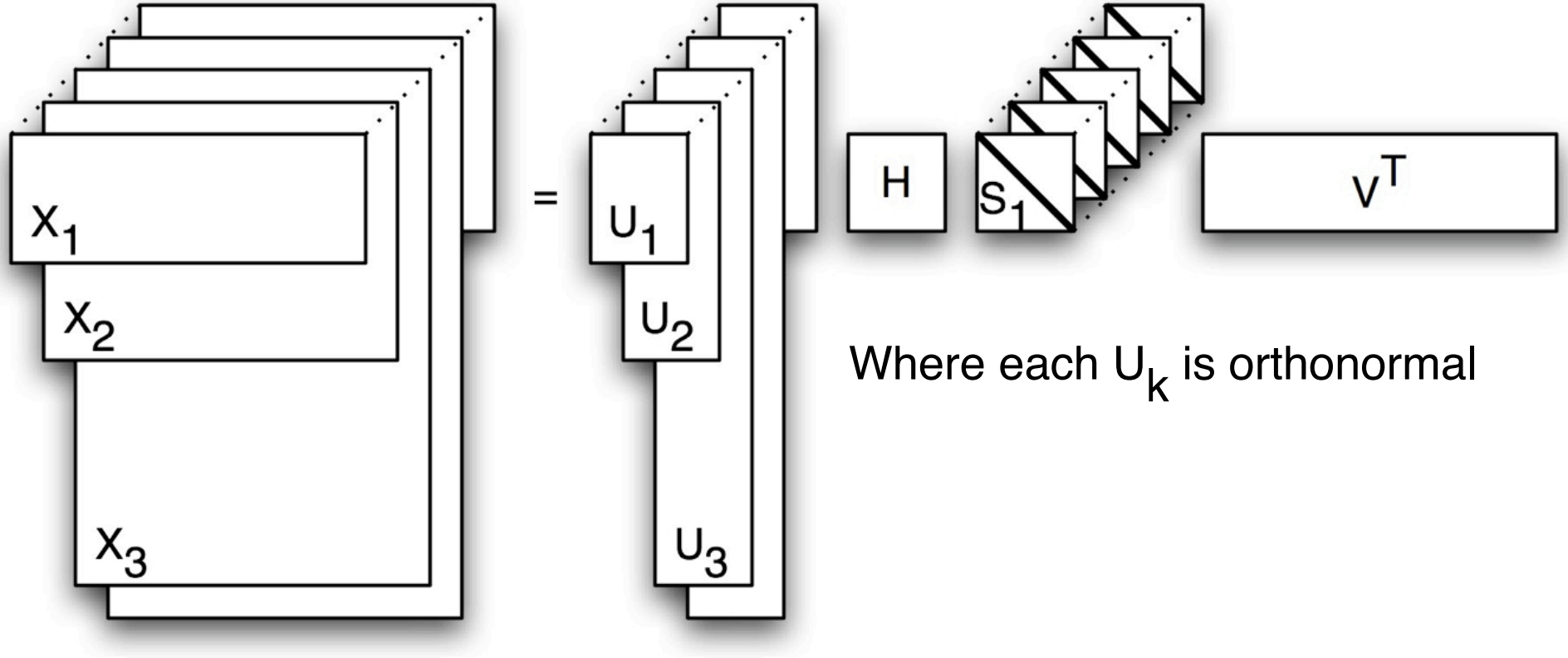
Term-by-verse matrix
for each language



Sandia National Laboratories

PARAFAC2

$$X_k \approx U_k H S_k V^T \quad \text{for } k = 1, \ldots, K$$



Where each U_k is orthonormal

Project documents of interest into subspace of corresponding U and compute cosine similarities

Clustering effectiveness (i.e., multilingual precision) improved from 26% to above 60%

Sandia National Laboratories

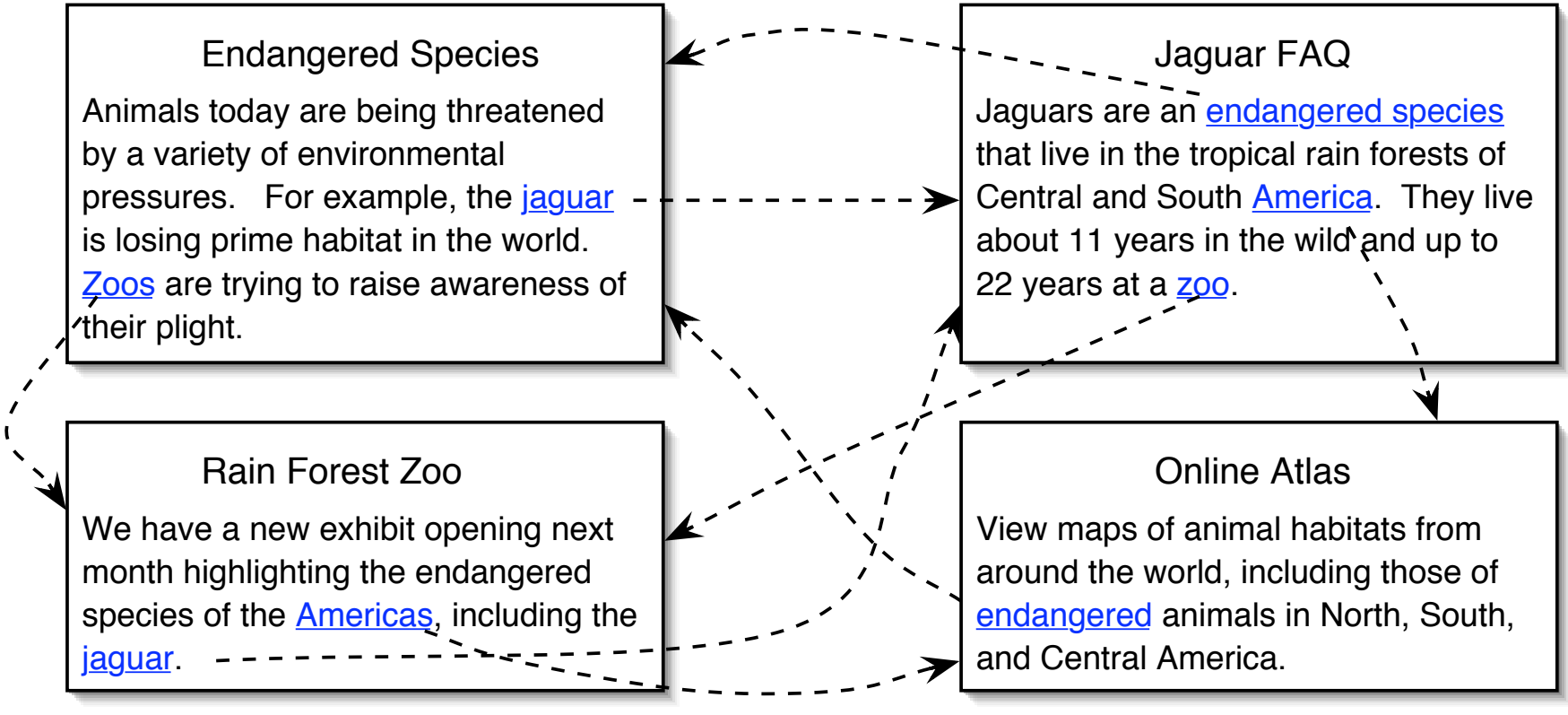
The TOPHITS model for higher-order web link analysis

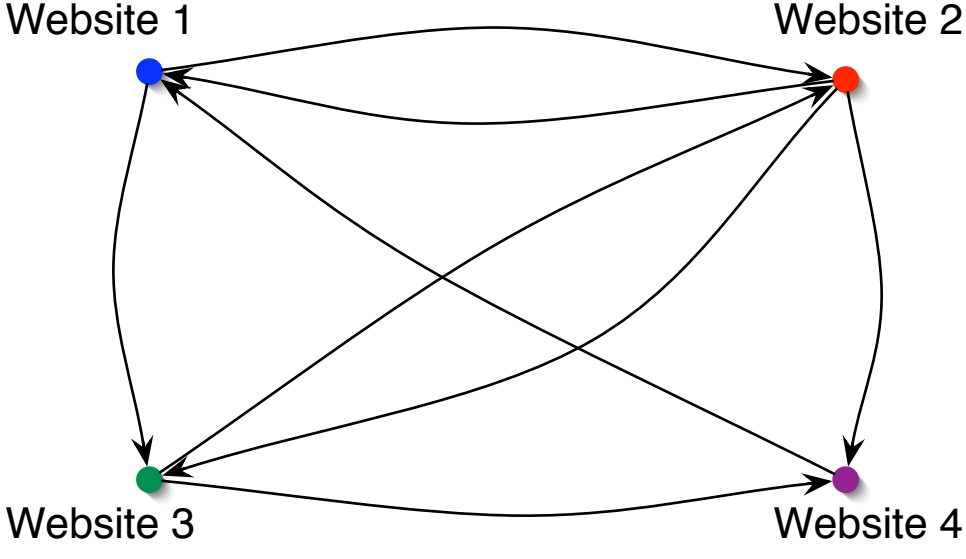
Tamara Kolda and Brett Bader
Sandia National Laboratories

Workshop on Link Analysis, Counterterrorism and Security
April, 2006

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.

The Web as a Graph





Hyperlinked structure of the web incorporates human perceptions of importance and relevance.

↓

Each hyperlink is like a vote.

Context information missing in this graph.

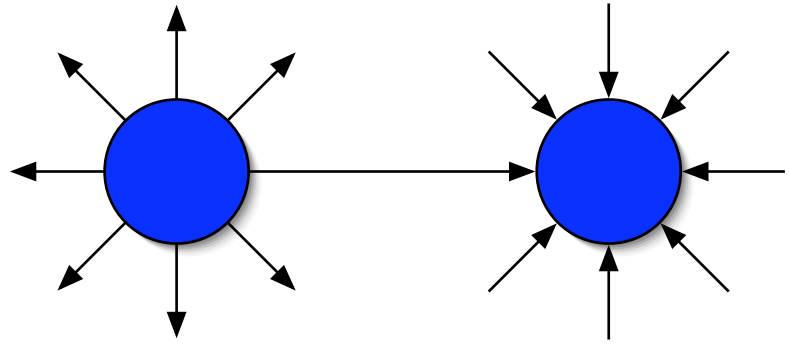
Sandia National Laboratories

HITS

Hypertext Induced Topic Search

- Developed by Kleinberg in 1998 (about the same time as PageRank)
- Variant used in the Teoma search engine*

Websites classified as 2 types



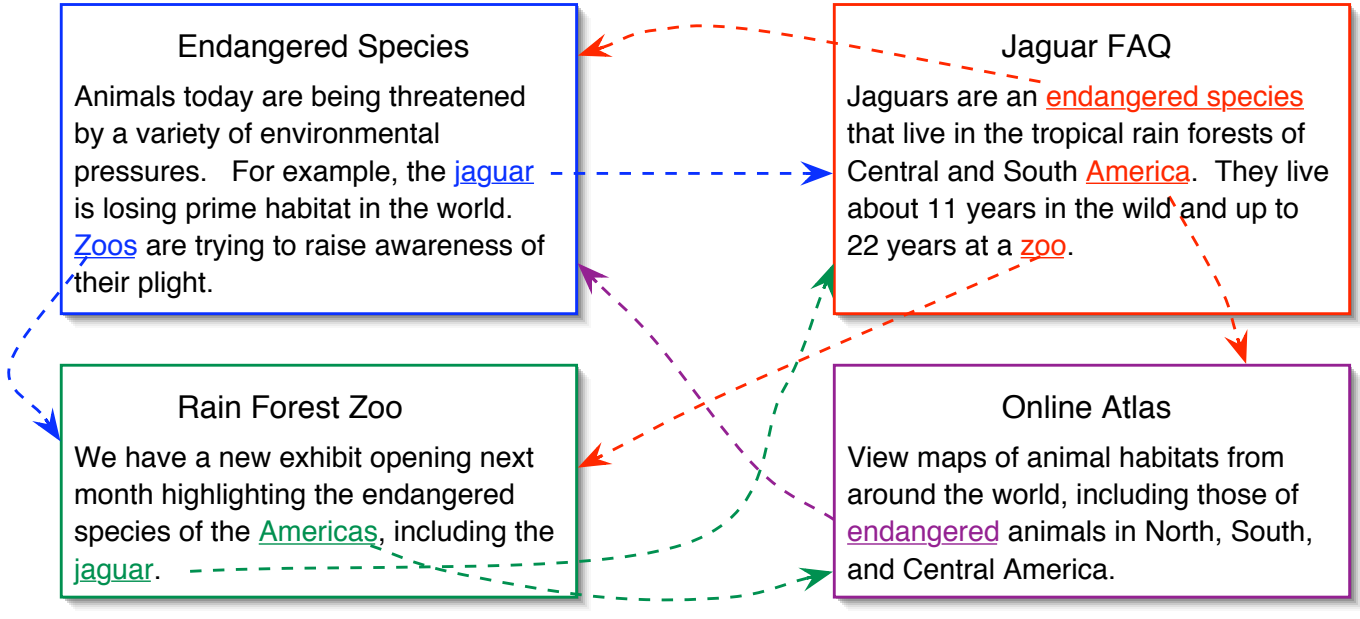
Hubs Authorities

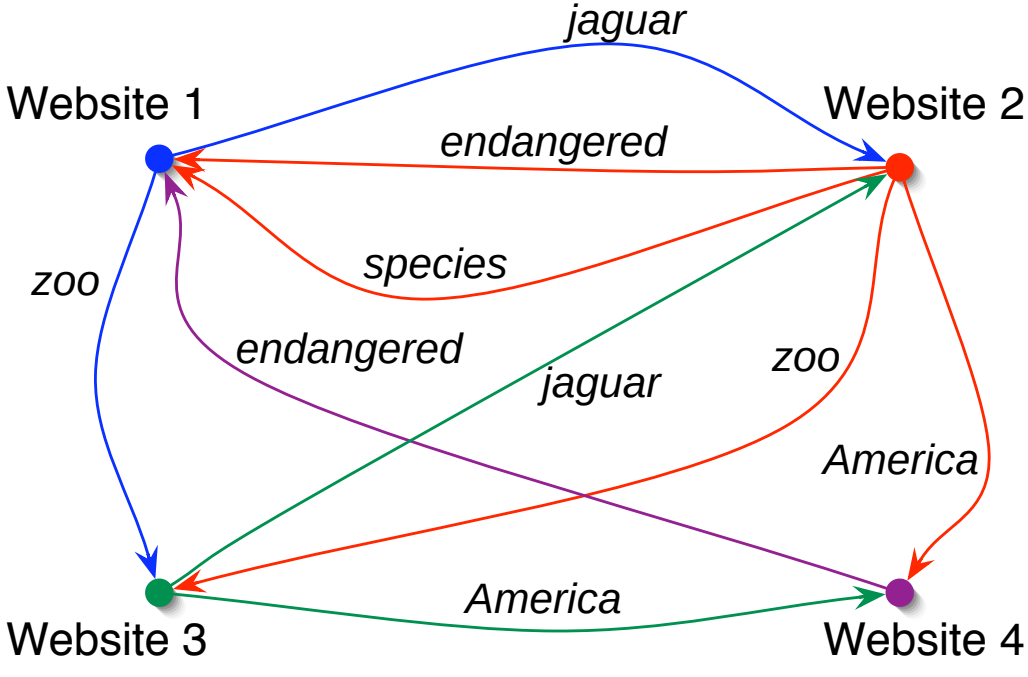
"Good hubs point to good authorities"

(mutually reinforcing relationship)

*(Langville and Meyer, 2005)

Web as a Semantic Graph

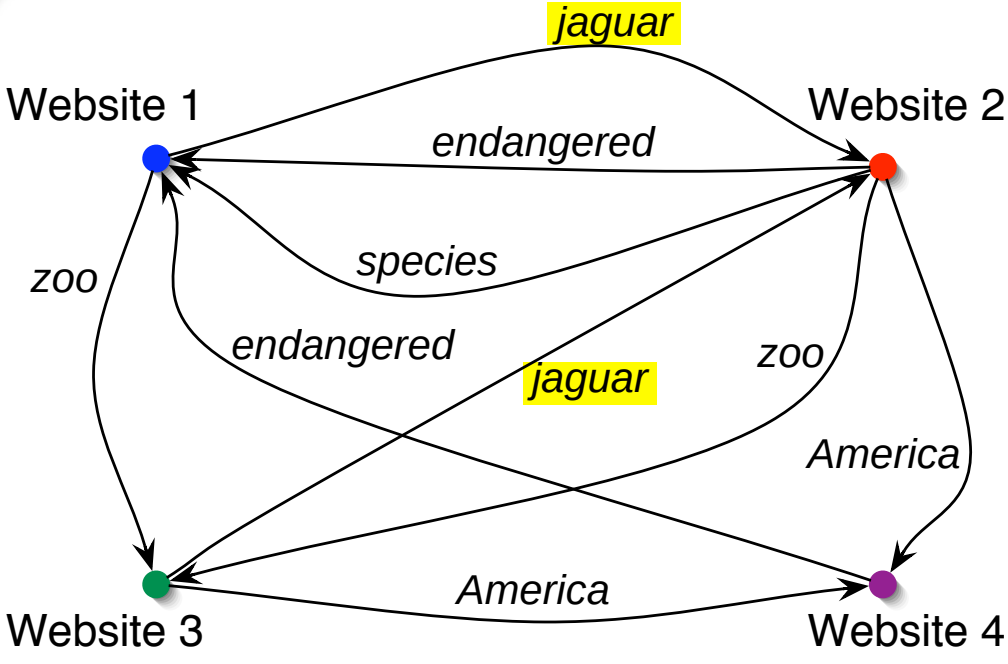




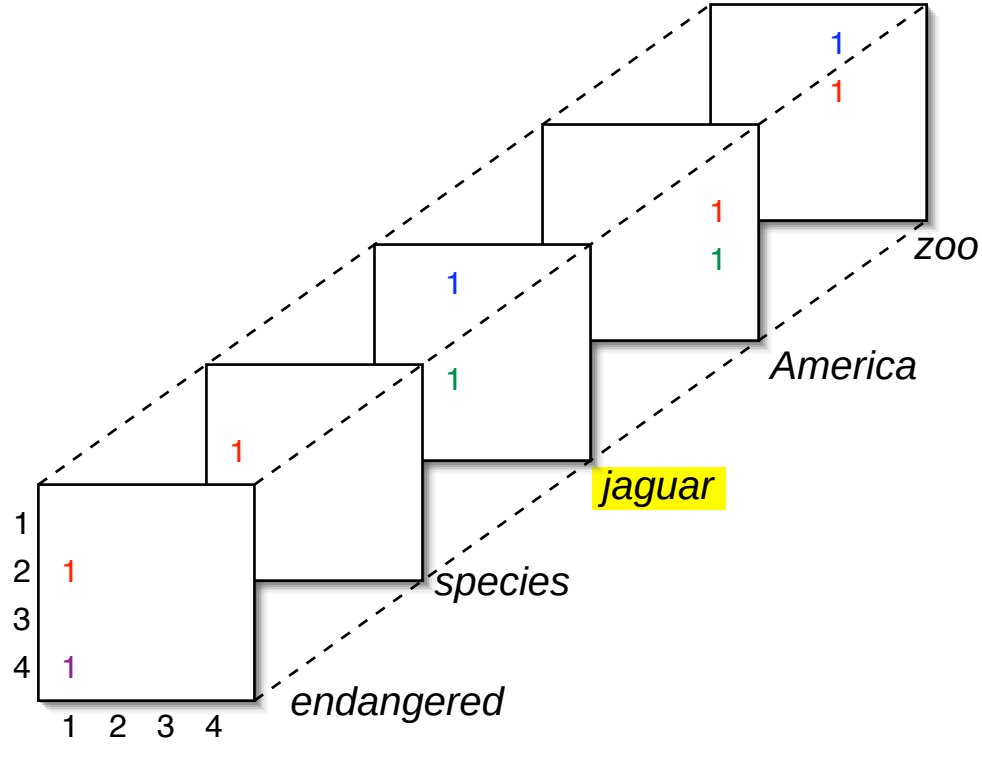
Edges labeled with anchor text

Sandia National Laboratories

Adjacency Tensor



Create an adjacency matrix for each edge type and store it as a slice in tensor.



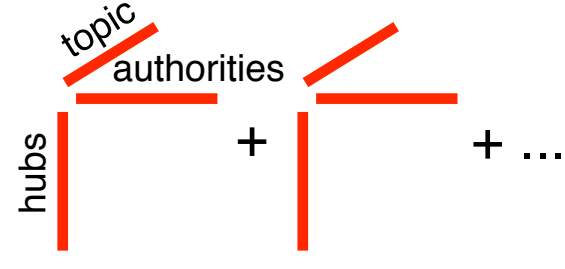
$$\mathcal{A}_{ijk} = \begin{cases} 1 & \text{if } i \to j \text{ with anchor text } k, \\ 0 & \text{otherwise.} \end{cases}$$

Perform multiway analysis on tensor

TOPHITS Results

Topics		Authorities	
SCORE	TERM	SCORE	HOST
1st Principal Factor			
0.23	java	0.86	java.sun.com
0.18	sun	0.38	developers.sun.com
0.17	platform	0.16	docs.sun.com
0.16	solaris	0.14	see.sun.com
0.16	developer	0.14	www.sun.com
0.15	edition	0.09	www.samag.com
0.15	download	0.07	developer.sun.com
0.14	info	0.06	sunsolve.sun.com
0.12	software	0.05	access1.sun.com
		0.05	iforce.sun.com
2nd Principal Factor			
0.20	no-anchor-text	0.99	www.lehigh.edu
0.16	faculty	0.06	www2.lehigh.edu
0.16	search	0.03	www.lehighalumni.com
0.16	news		
0.16	libraries		
0.16	computing		
0.12	lehigh		
3rd Principal Factor			
0.15	no-anchor-text	0.97	www.ibm.com
0.15	ibm	0.18	www.alphaworks.ibm.com
0.12	services	0.07	www-128.ibm.com
0.12	websphere	0.05	www.developer.ibm.com
0.12	web	0.02	www.redbooks.ibm.com
0.11	developerworks	0.01	www.research.ibm.com
0.11	linux		
0.11	resources		
0.11	technologies		
0.10	downloads		

Topics		Authorities	
SCORE	TERM	SCORE	HOST
4th Principal Factor			
0.26	information	0.87	www.pueblo.gsa.gov
0.24	federal	0.24	www.irs.gov
0.23	citizen	0.23	www.whitehouse.gov
0.22	other	0.19	travel.state.gov
0.19	center	0.18	www.gsa.gov
0.19	languages	0.09	www.consumer.gov
0.15	u.s	0.09	www.kids.gov
0.15	publications	0.07	www.ssa.gov
0.14	consumer	0.05	www.forms.gov
0.13	free	0.04	www.govbenefits.gov
6th Principal Factor			
0.26	president	0.87	www.whitehouse.gov
0.25	no-anchor-text	0.18	www.irs.gov
0.25	bush	0.16	travel.state.gov
0.25	welcome	0.10	www.gsa.gov
0.17	white	0.08	www.ssa.gov
0.16	u.s		
0.15	house		
0.13	budget		
0.13	presidents		
0.11	office		



Sandia National Laboratories

More Information

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

- MATLAB Tensor Toolbox version 2.2:
- http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox