



# Large Graph Mining: Power Tools and a Practitioner's guide

Task 5: Graphs over time & tensors

*Faloutsos, Miller, Tsourakakis*

CMU



# Outline

- Introduction – Motivation
- Task 1: Node importance
- Task 2: Community detection
- Task 3: Recommendations
- Task 4: Connection sub-graphs
- ➔ • **Task 5: Mining graphs over time**
- Task 6: Virus/influence propagation
- Task 7: Spectral graph theory
- Task 8: Tera/peta graph mining: hadoop
- Observations – patterns of real graphs
- Conclusions



## Thanks to

- Tamara Kolda (Sandia)



for the foils on tensor  
definitions, and on TOPHITS



## Detailed outline

- Motivation
- Definitions: PARAFAC and Tucker
- Case study: web mining



# Examples of Matrices:

## Authors and terms

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...



# Motivation: Why tensors?

- Q: what is a tensor?



# Motivation: Why tensors?

- A: N-D generalization of matrix:

KDD'09	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...



# Motivation: Why tensors?

- A: N-D generalization of matrix:

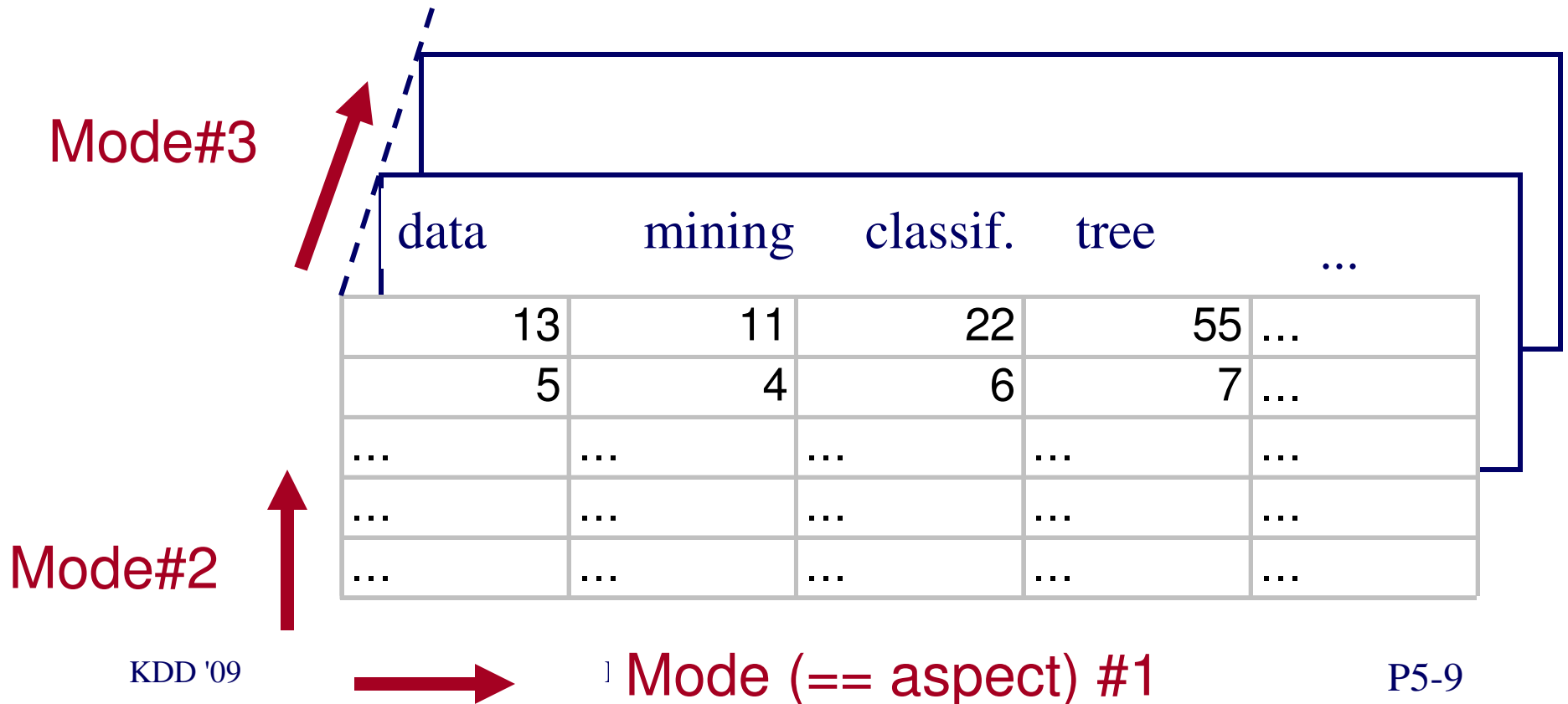
KDD'07					
KDD'08					
KDD'09	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...





# Tensors are useful for 3 or more modes

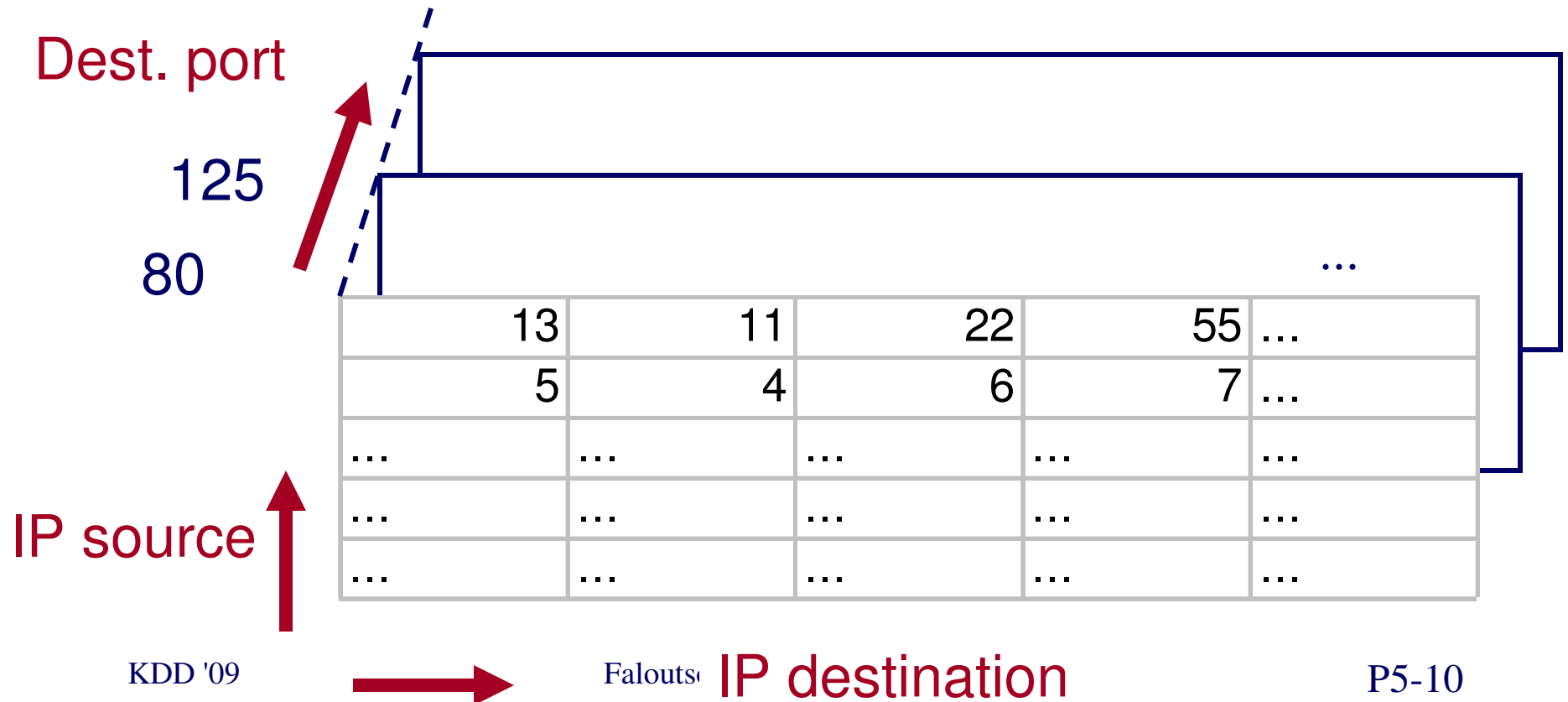
Terminology: ‘mode’ (or ‘aspect’):





## Notice

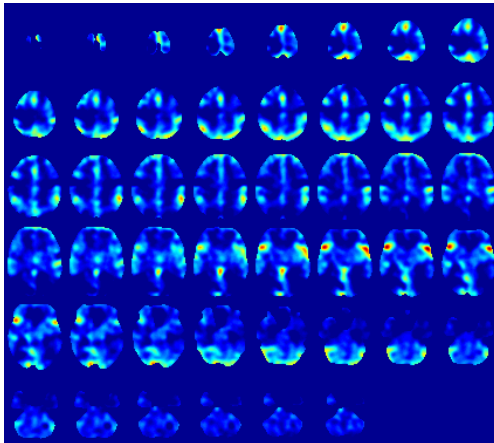
- 3<sup>rd</sup> mode does not need to be time
- we can have more than 3 modes





## Notice

- 3<sup>rd</sup> mode does not need to be time
- we can have more than 3 modes
  - Eg, fMRI: x,y,z, time, person-id, task-id



From DENLAB, Temple U.  
(Prof. V. Megalooikonomou +)

<http://denlab.temple.edu/bidms/cgi-bin/browse.cgi>



# Motivating Applications

- Why tensors are useful?
  - web mining (TOPHITS)
  - environmental sensors
  - Intrusion detection (src, dst, time, dest-port)
  - Social networks (src, dst, time, type-of-contact)
  - face recognition
  - etc ...



## Detailed outline

- Motivation
- ➔ • Definitions: PARAFAC and Tucker
- Case study: web mining



# Tensor basics

- Multi-mode extensions of SVD – recall that:



## Reminder: SVD

$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$

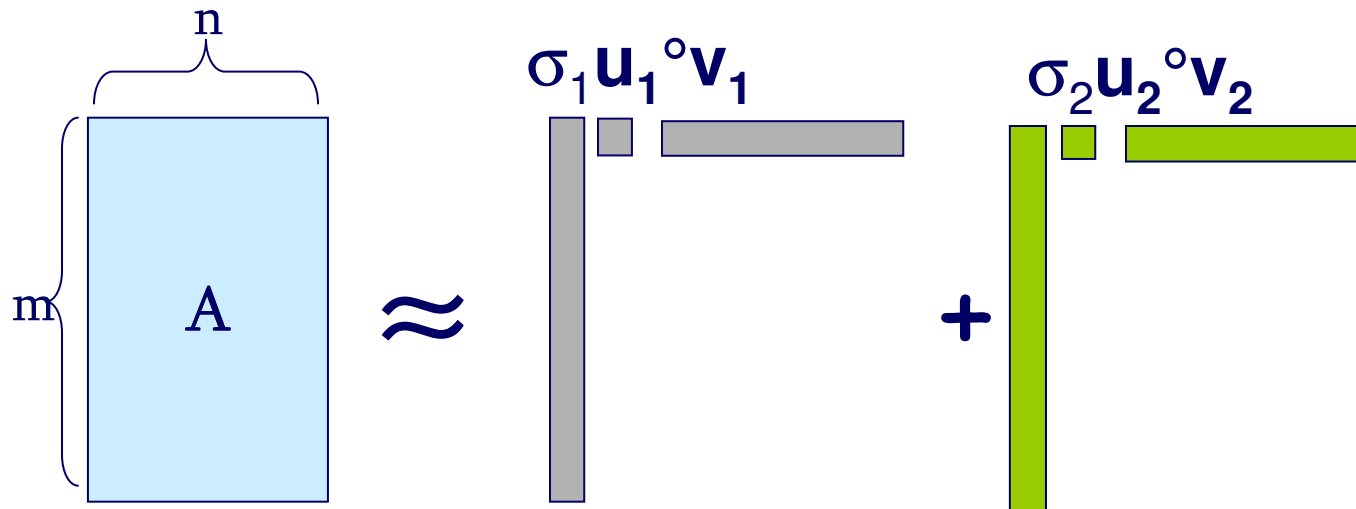
The diagram illustrates the SVD decomposition of matrix  $\mathbf{A}$  (size  $m \times n$ ) into three matrices:  $\mathbf{U}$  (size  $m \times k$ ),  $\mathbf{\Sigma}$  (size  $k \times k$ ), and  $\mathbf{V}^T$  (size  $k \times n$ ). Matrix  $\mathbf{U}$  is represented by a vertical rectangle with a green stripe. Matrix  $\mathbf{\Sigma}$  is a small square with a green stripe. Matrix  $\mathbf{V}^T$  is a horizontal rectangle with a green stripe. The dimensions  $m$ ,  $n$ , and  $k$  are indicated by brackets.

– Best rank- $k$  approximation in L2



## Reminder: SVD

$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$

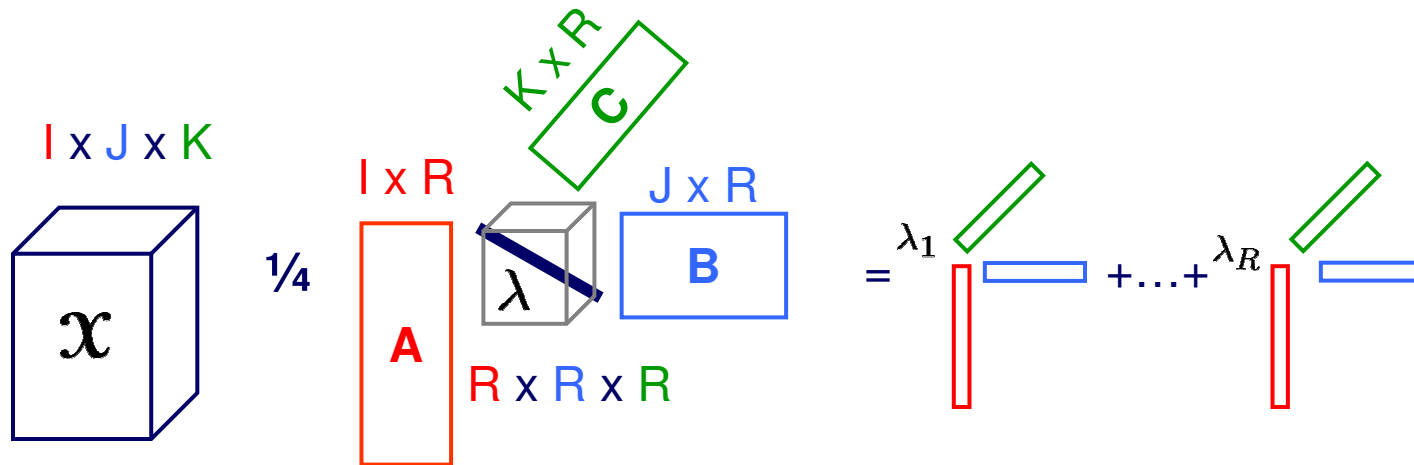


– Best rank- $k$  approximation in L2





# Goal: extension to $\geq 3$ modes



$$\mathcal{X} \approx [\lambda; A, B, C] = \sum_r \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$



## Main points:

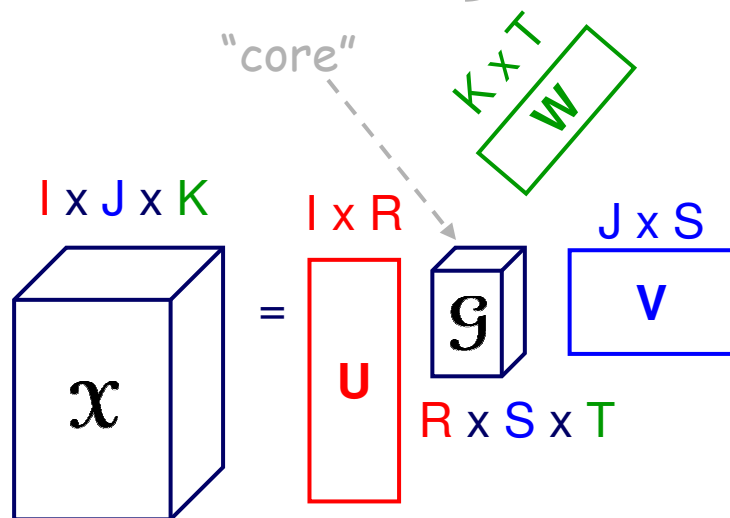
- 2 major types of tensor decompositions: PARAFAC and Tucker
- both can be solved with ‘‘alternating least squares’’ (ALS)



# Specially Structured Tensors

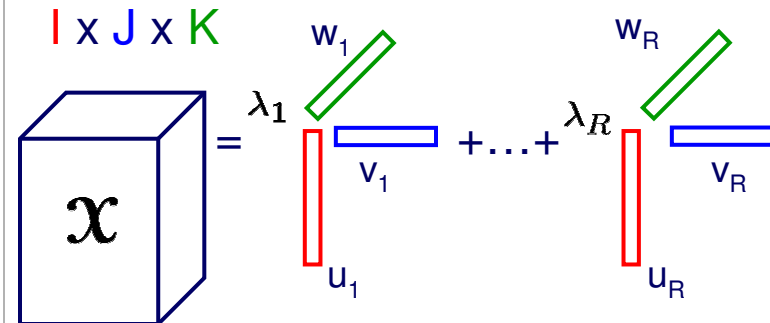
## • Tucker Tensor

$$\begin{aligned}
 \mathcal{X} &= \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} \\
 &= \sum_r \sum_s \sum_t g_{rst} \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t \\
 &\equiv [\![\mathcal{G} ; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!] \quad \left. \vphantom{\sum_r \sum_s \sum_t} \right\} \text{Our Notation}
 \end{aligned}$$



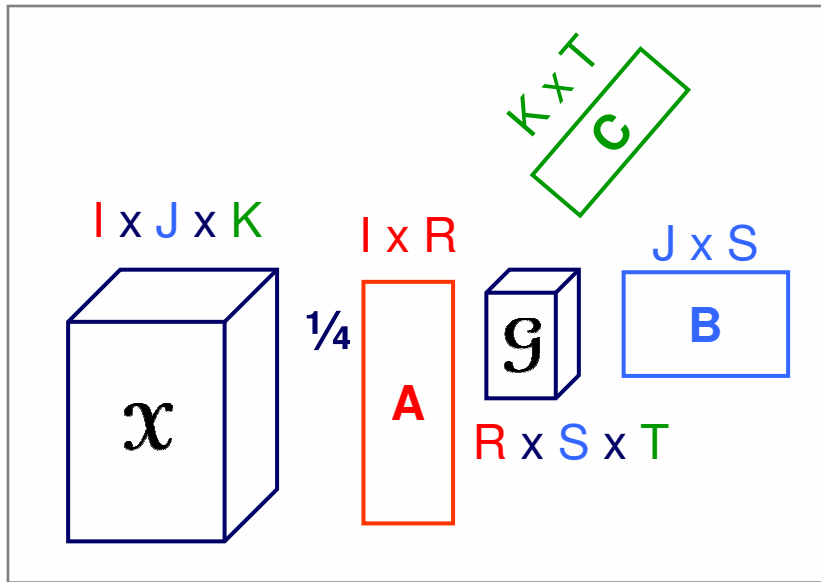
## • Kruskal Tensor

$$\begin{aligned}
 \mathcal{X} &= \sum_r \lambda_r \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \\
 &\equiv [\![\lambda ; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!] \quad \left. \vphantom{\sum_r} \right\} \text{Our Notation}
 \end{aligned}$$





# Tucker Decomposition - intuition



- author x keyword x conference
- $\mathbf{A}$ : author x author-group
- $\mathbf{B}$ : keyword x keyword-group
- $\mathbf{C}$ : conf. x conf-group
- $\mathbf{G}$ : how groups relate to each other



# Intuition behind core tensor

- 2-d case: co-clustering
- [Dhillon et al. Information-Theoretic Co-clustering, KDD'03]



$$\begin{matrix} & \overline{\hspace{1cm}} & \overline{\hspace{1cm}} \\ & n & \\ m & \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \end{matrix}$$

eg, terms x documents

$$\begin{matrix} & k & & l & & n \\ m & \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} & k & \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} & l & \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} & = & \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ \hline .036 & .036 & .028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix} \end{matrix}$$



med. doc      cs doc

term group x  
doc. group

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

med. terms

cs terms

common terms

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix}$$

$$\begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}$$

$$\begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} =$$

$$\begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & .028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

doc x  
doc group

term x  
term-group



## Tensor tools - summary

- Two main tools
  - PARAFAC
  - Tucker
- Both find row-, column-, tube-groups
  - but in PARAFAC the three groups are identical
- ( To solve: Alternating Least Squares )





## Detailed outline

- Motivation
- Definitions: PARAFAC and Tucker
- ➔ • Case study: web mining



# Web graph mining

- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (**TOPHITS**)

Google Web Images Video News Maps more »

tensor Search Advanced Search Preferences

Turn OFF Personalized Search (Beta) for these results »

Web Personalized Results 1 - 10 of about 12,800,000 for tensor [definition]. (0.31 seconds)

**Tensor** - Wikipedia, the free encyclopedia  
Examples of physical **tensors** are the energy-momentum **tensor**, the inertia **tensor** ...  
Tensorial 3.0 Tensorial is a general purpose **tensor** calculus package for ...  
[en.wikipedia.org/wiki/Tensor](http://en.wikipedia.org/wiki/Tensor) - 55k - [Cached](#) - [Similar pages](#)

**Tensor product** - Wikipedia, the free encyclopedia  
There is a general formula for the product of two (or more) **tensors**, as ... The **tensor** product inherits all the indices of its factors. ...  
[en.wikipedia.org/wiki/Tensor\\_product](http://en.wikipedia.org/wiki/Tensor_product) - 41k - [Cached](#) - [Similar pages](#)

**Tensor Trucks**  
Manufacturer of skateboard trucks. Check out team members, videos and apparel.  
[www.tensortrucks.com/](http://www.tensortrucks.com/) - 3k - [Cached](#) - [Similar pages](#)

**Time and Attendance & Access Control through Smart Cards ...**  
**Tensor** manufacture and supply Smart Card and Biometric Time and Attendance & Access Control Software and Systems.  
[www.tensor.co.uk/](http://www.tensor.co.uk/) - 7k - [Cached](#) - [Similar pages](#)

**Free Textbook Tensor Calculus and Continuum Mechanics**  
A free downloadable textbook on introductory **tensor** analysis and continuum

Sponsored Links

**Tensor**  
Bargain Prices. Smart Deals. Save on **Tensor**!  
[Shopzilla.com](http://Shopzilla.com)

**Wire Tensioners**  
For coil and motor winding machines Mechanical or electronic tensioners  
[www.digimotor.com](http://www.digimotor.com)

**Tensor**  
Looking for **Tensor**? Find exactly what you want today  
[www.eBay.com](http://www.eBay.com)

**Tensor**  
Shop For **Tensor** Here With The Convenience Of OneCart™!  
[SHOP.COM](http://SHOP.COM)

YAHOO! SEARCH

Web Images Video Local Shopping more »

tensor Search

Search Results 1 - 10 of about 2,870,000 for tensor - 0.74 sec. (About this page)

Also try: [tensor lamps](#), [tensor lighting](#), [tensor corporation](#), [tensor product](#) More...

Sponsored Results

- Tensor Skateboard Trucks**  
[www.AllegroMedical.com](http://www.AllegroMedical.com) - Great Selection and Fast Shipping Order Online Today and Save.
- Purchase Tensor Bandages at HCD**  
[www.homecaredelivered.com](http://www.homecaredelivered.com) - Save on our full line of wound care supplies.

- Tensor - from MathWorld**  
An  $n$ -rank **tensor** in  $m$ -dimensional space is a mathematical object that has  $n$  ... Each index of a **tensor** ranges over the number of dimensions of space. ...  
[mathworld.wolfram.com/Tensor.html](http://mathworld.wolfram.com/Tensor.html) - [More from this site](#)
- Tensor - Wikipedia, the free encyclopedia**  
The term **tensor** has slightly different meanings in mathematics and physics. ... algebra and differential geometry, a **tensor** is a multilinear function. ...  
Quick Links: [Importance and applications](#) - [History](#) - [The choice of approach](#)  
[en.wikipedia.org/wiki/Tensor](http://en.wikipedia.org/wiki/Tensor) - 50k - [Cached](#) - [More from this site](#)

Sponsor Results

**Tensor**  
Find Deals on **Tensor** and other Sporting Equipment at DealTime....  
[www.dealtime.com](http://www.dealtime.com)

**Tensor Compare Prices**  
Find Bargains on **Tensor** at thousands of trusted online stores. Get...  
[www.bizrate.com](http://www.bizrate.com)

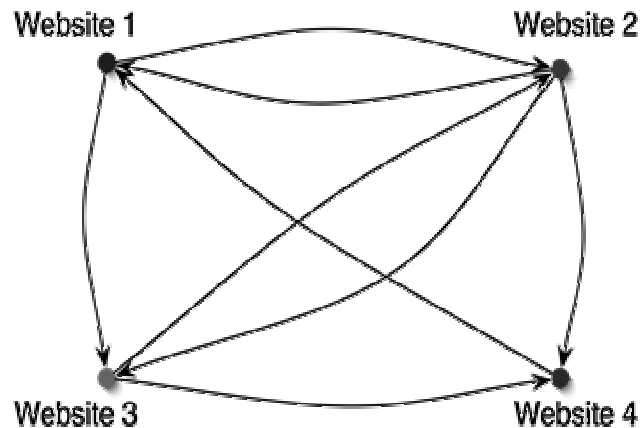
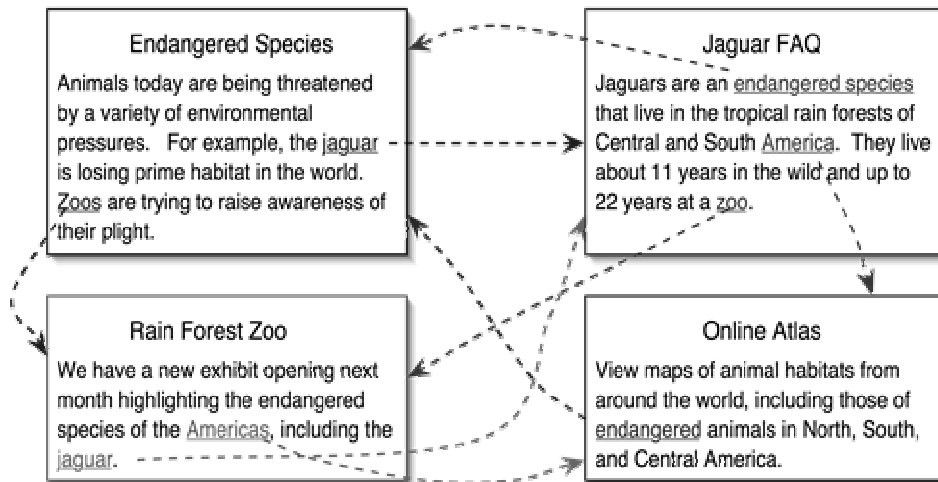
**Tensor**  
We are writing an on-line e-book with code: "Pseudocolor in Pure...  
[www.youvan.com](http://www.youvan.com)

**Tensor at Shopping.com**  
Find, compare and buy products in categories ranging from sports...  
[www.shopping.com](http://www.shopping.com)

**Tensor**  
Shop eBay for anything and



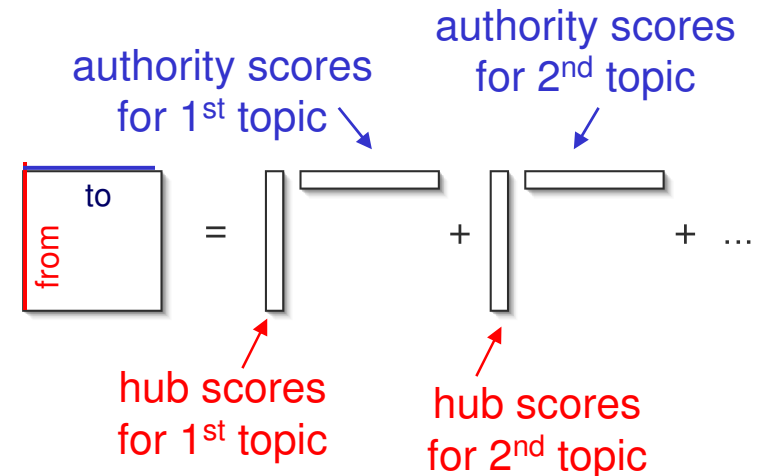
# Kleinberg's Hubs and Authorities (the HITS method)



Sparse adjacency matrix and its SVD:

$$x_{ij} = \begin{cases} 1 & \text{if page } i \text{ links to page } j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{X} \approx \sum_r \sigma_r \mathbf{h}_r \circ \mathbf{a}_r$$





# HITS Authorities on Sample Data

1st Principal Factor	
.97	www.ibm.com
.24	www.alphaw.com
.08	www-128.ibm.com
.05	www.develop.com
.02	www.research.com
.01	www.redbook.com
.01	news.com.com

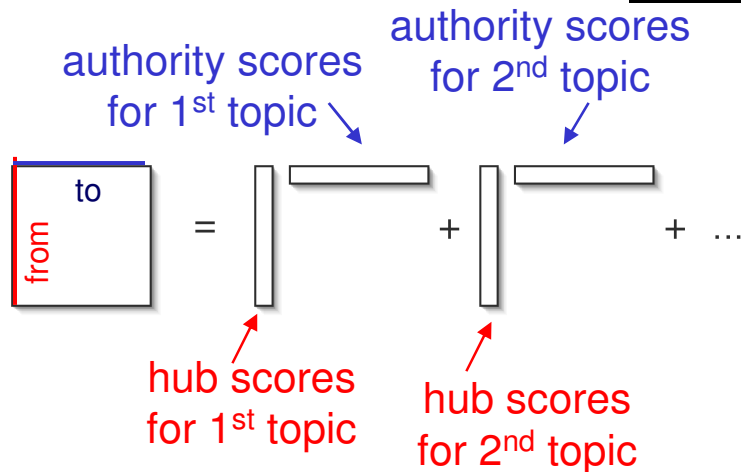
2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh.edu
.06	www.lehigh.edu
.06	www.lehigh.edu
.02	www.bethleham.edu
.02	www.adobe.com
.02	lewisweb.cc
.02	www.leo.lehigh.edu
.02	www.distance.edu
.02	fp1.cc.lehigh.edu

3rd Principal Factor	
.75	java.sun.com
.38	www.sun.com
.36	developers.sun.com
.24	see.sun.com
.16	www.samag.com
.13	docs.sun.com
.12	blogs.sun.com
.08	sunsolve.sun.com
.08	www.sun-catalog.com
.08	news.com.com

4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state.gov
.22	www.gsa.gov
.20	www.ssa.gov
.16	www.census.gov
.14	www.govbeat.com
.13	www.kids.gov
.13	www.usdoj.gov

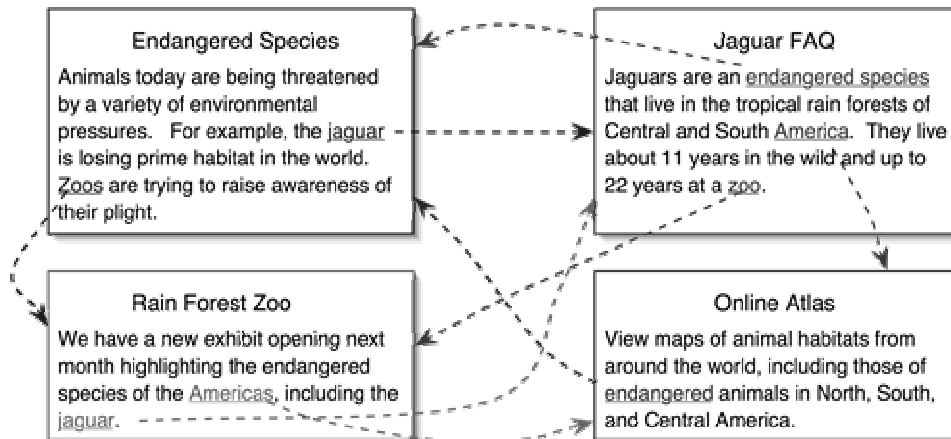
We started our crawl from  
<http://www-neos.mcs.anl.gov/neos>,  
and crawled 4700 pages,  
resulting in 560  
cross-linked hosts.

6th Principal Factor	
.97	mathpost.asu.edu
.18	math.la.asu.edu
.17	www.asu.edu
.04	www.act.org
.03	www.eas.asu.edu
.02	archives.math.utk.edu
.02	www.geom.uiuc.edu
.02	www.fulton.asu.edu
.02	www.amstat.org
.02	www.maa.org



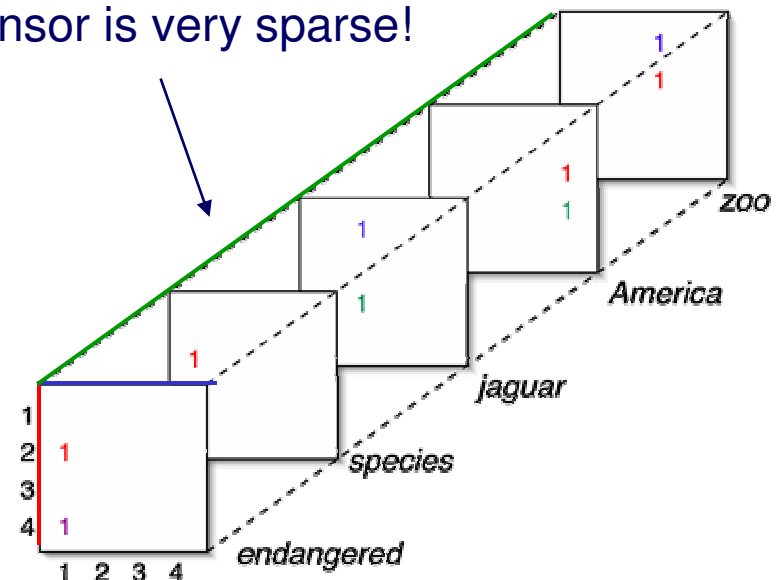
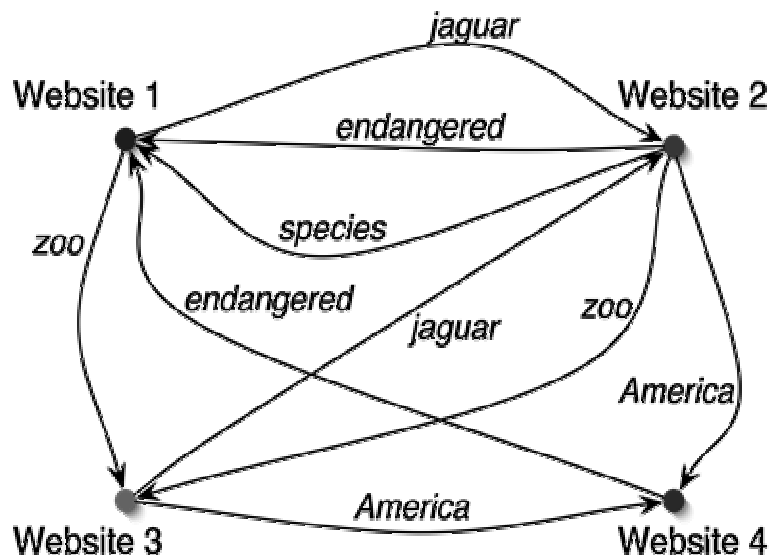


# Three-Dimensional View of the Web



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

Observe that this tensor is very sparse!





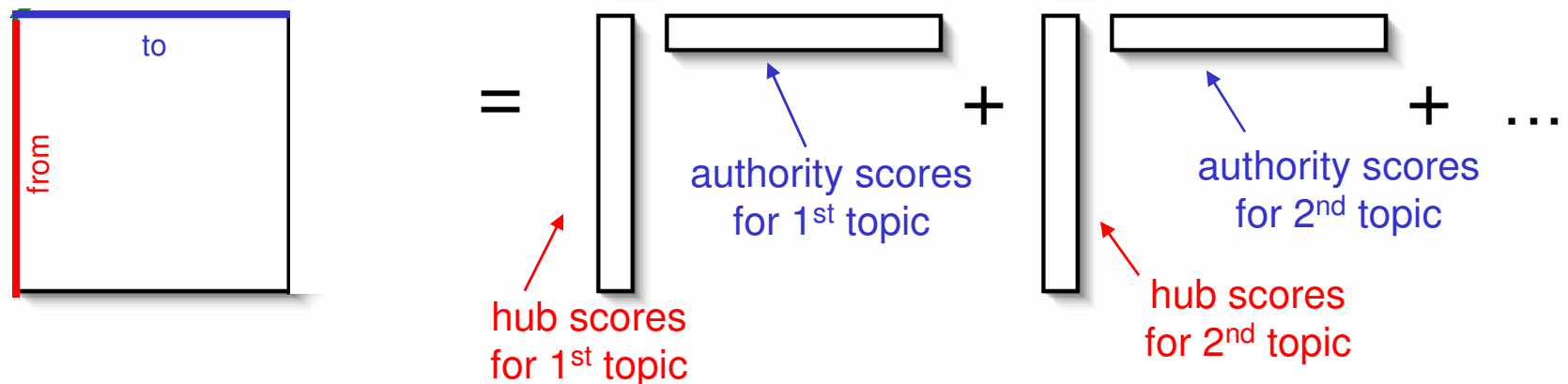




# Topical HITS (TOPHITS)

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

$$\mathbf{x} \approx \sum_{r=1}^R \lambda_r \mathbf{h}_r \circ \mathbf{a}_r$$



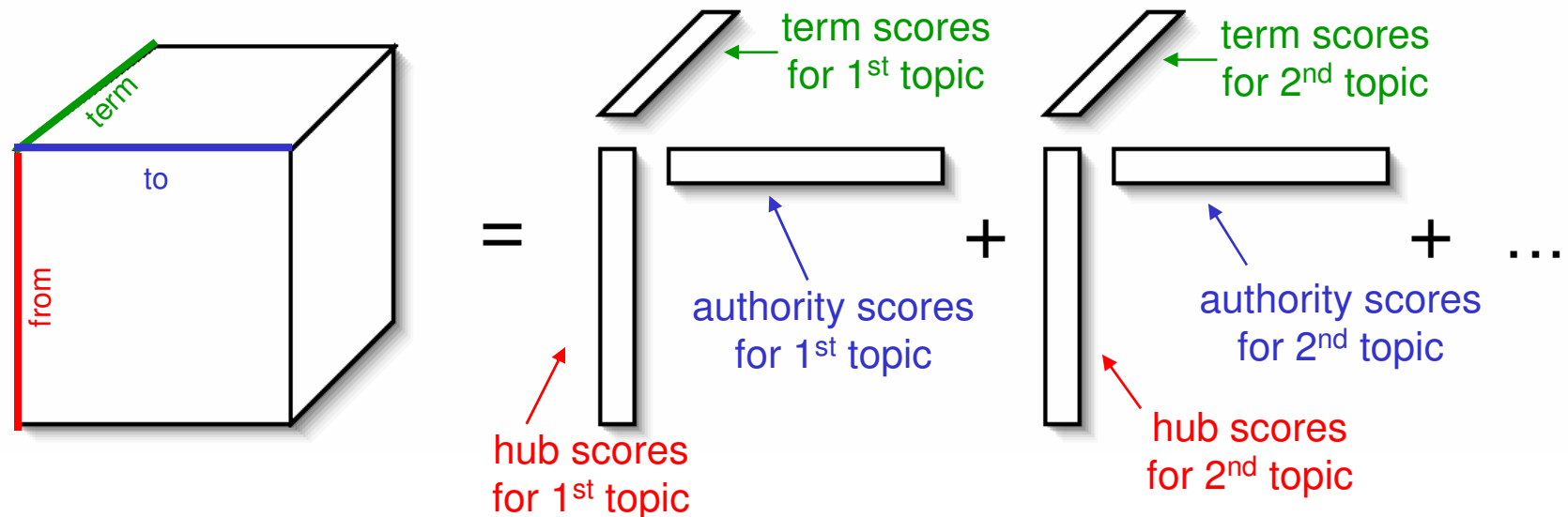




# Topical HITS (TOPHITS)

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

$$\mathbf{x} \approx \sum_{r=1}^R \lambda_r \mathbf{h}_r \circ \mathbf{a}_r \circ \mathbf{t}_r$$





CMU SCS

# TOPHITS Terms & Authorities

## on Sample Data

1st Principal Factor			
.23	JAVA	.86	java.sun.com
.18	SUN	.98	developers.sun.com

2nd Principal Factor			
.17	PLATF	.20	NO-READABLE-TEXT
.16	SOLAR	.99	www.lehigh.edu

3rd Principal Factor			
.16	DEVEL	.15	NO-READABLE-TEXT
.15	EDITIO	.97	www.ibm.com

4th Principal Factor			
.15	DOWN	.15	IBM
.14	INFO	.18	www.alphaworks.ibm.com

6th Principal Factor			
.11	DEVEL	.26	INFORMATION
.11	LINUX	.87	www.pueblo.gsa.gov

12th Principal Factor			
.11	RESOL	.26	PRESIDENT
.11	TECHN	.87	www.whitehouse.gov

13th Principal Factor			
.10	DOWN	.25	WELC
.15	PUBLIC	.35	www.palisade.com

16th Principal Factor			
.13	FREE	.46	ADOBE
.15	HOUS	.99	www.adobe.com

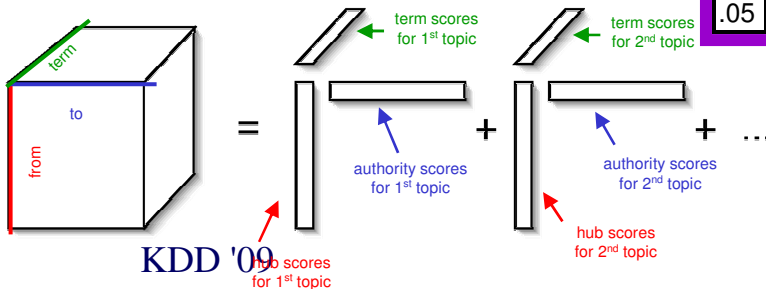
19th Principal Factor			
.05	ILOG	.22	TAX
.05	DOWN	.73	www.irs.gov

TOPHITS uses 3D analysis to find the dominant groupings of web pages and terms.

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

$w_k$  = # unique links using term  $k$

Tensor PARAFAC





## Conclusions

- Real data are often in high dimensions with multiple aspects (modes)
- Tensors provide elegant theory and algorithms
  - PARAFAC and Tucker: discover groups



## References

- T. G. Kolda, B. W. Bader and J. P. Kenny.  
*Higher-Order Web Link Analysis Using Multilinear Algebra*. In: ICDM 2005, Pages 242-249, November 2005.
- Jimeng Sun, Spiros Papadimitriou, Philip Yu.  
*Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams*, Proc. of the Int. Conf. on Data Mining (ICDM), Hong Kong, China, Dec 2006



## Resources

- See tutorial on tensors, KDD'07 (w/ Tamara Kolda and Jimeng Sun):

[www.cs.cmu.edu/~christos/TALKS/KDD-07-tutorial](http://www.cs.cmu.edu/~christos/TALKS/KDD-07-tutorial)



## Tensor tools - resources



- Toolbox: from Tamara Kolda:  
[csmr.ca.sandia.gov/~tgkolda/TensorToolbox](http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox)

- T. G. Kolda and B. W. Bader. ***Tensor Decompositions and Applications***. SIAM Review, Volume 51, Number 3, September 2009  
[csmr.ca.sandia.gov/~tgkolda/pubs/bibtgkfiles/TensorReview-preprint.pdf](http://csmr.ca.sandia.gov/~tgkolda/pubs/bibtgkfiles/TensorReview-preprint.pdf)
- T. Kolda and J. Sun: Scalable Tensor Decomposition for Multi-Aspect Data Mining (ICDM 2008)