CS665: Advanced Data Mining

Lecture#18: Tensor

U Kang
KAIST

Most of slides by

- Dr. Tamara Kolda (Sandia N.L.)
- http://csmr.ca.sandia.gov/~tgkolda

- Dr. Jimeng Sun (Georgia Tech)
- http://www.sunlab.org

3h tutorial: www.cs.cmu.edu/~christos/TALKS/SDM-tut-07/

Outline

- ☐ **Motivation - Definitions**
- ☐ Tensor tools
- ☐ Case Studies
- ☐ Conclusion

Motivation 0: Why "matrix"?

- Why matrices are important?

Examples of Matrices:
Graph - social network

	John	Peter	Mary	Nick	...
John	0	11	22	55	...
Peter	5	0	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices:
cloud of n-d points

	chol#	blood#	age	..	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices: Market basket

- **market basket** as in Association Rules

	milk	bread	choc.	wine	…
John	13	11	22	55	…
Peter	5	4	6	7	…
Mary	…	…	…	…	…
Nick	…	…	…	…	…
…	…	…	…	…	…

Examples of Matrices: Documents and terms

	data	mining	classif.	tree	...
Paper#1	13	11	22	55	...
Paper#2	5	4	6	7	...
Paper#3	...	...	...	...	...
Paper#4	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices: Authors and terms

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

Examples of Matrices: sensor-ids and time-ticks

	temp1	temp2	humid.	pressure	...
t1	13	11	22	55	...
t2	5	4	6	7	...
t3	...	...	...	...	...
t4	...	...	...	...	...
...	...	...	...	...	...

Motivation: Why tensors?

- Q: what is a tensor?

Motivation 2: Why tensor?

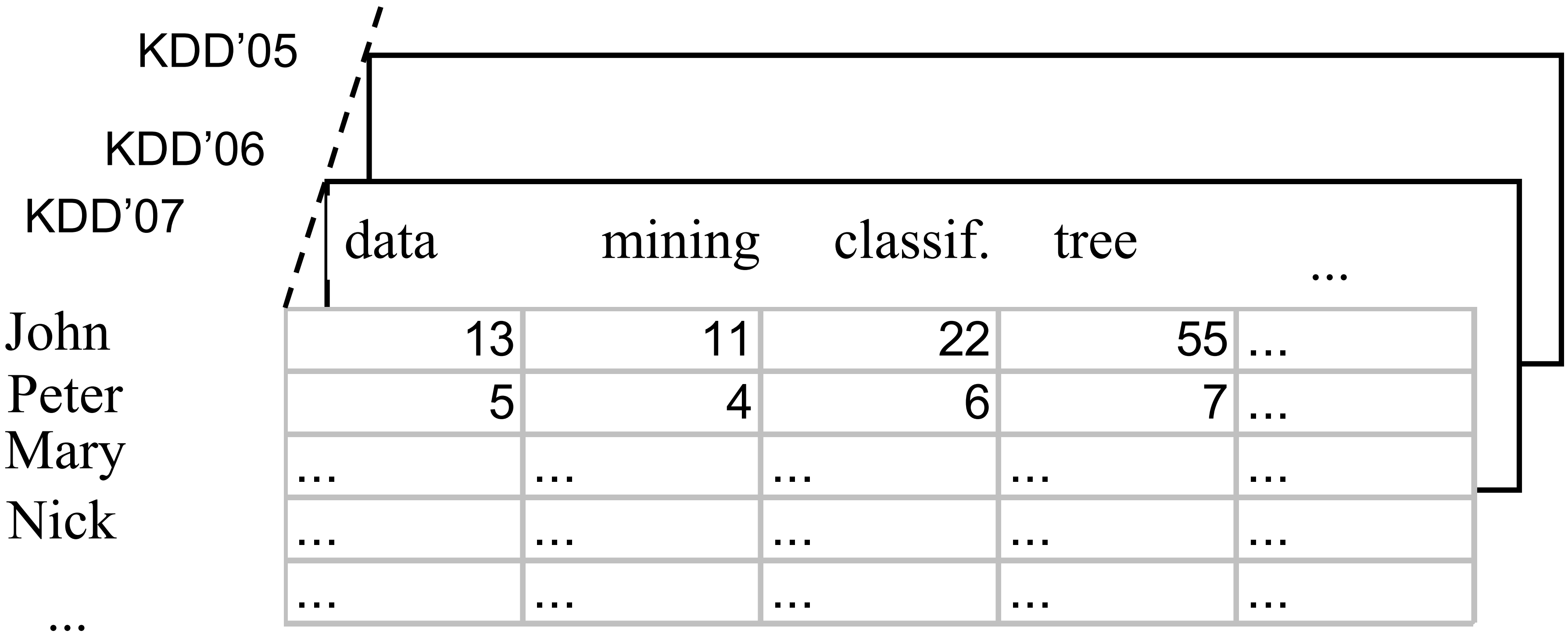
- A: N-D generalization of matrix:

KDD'07

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

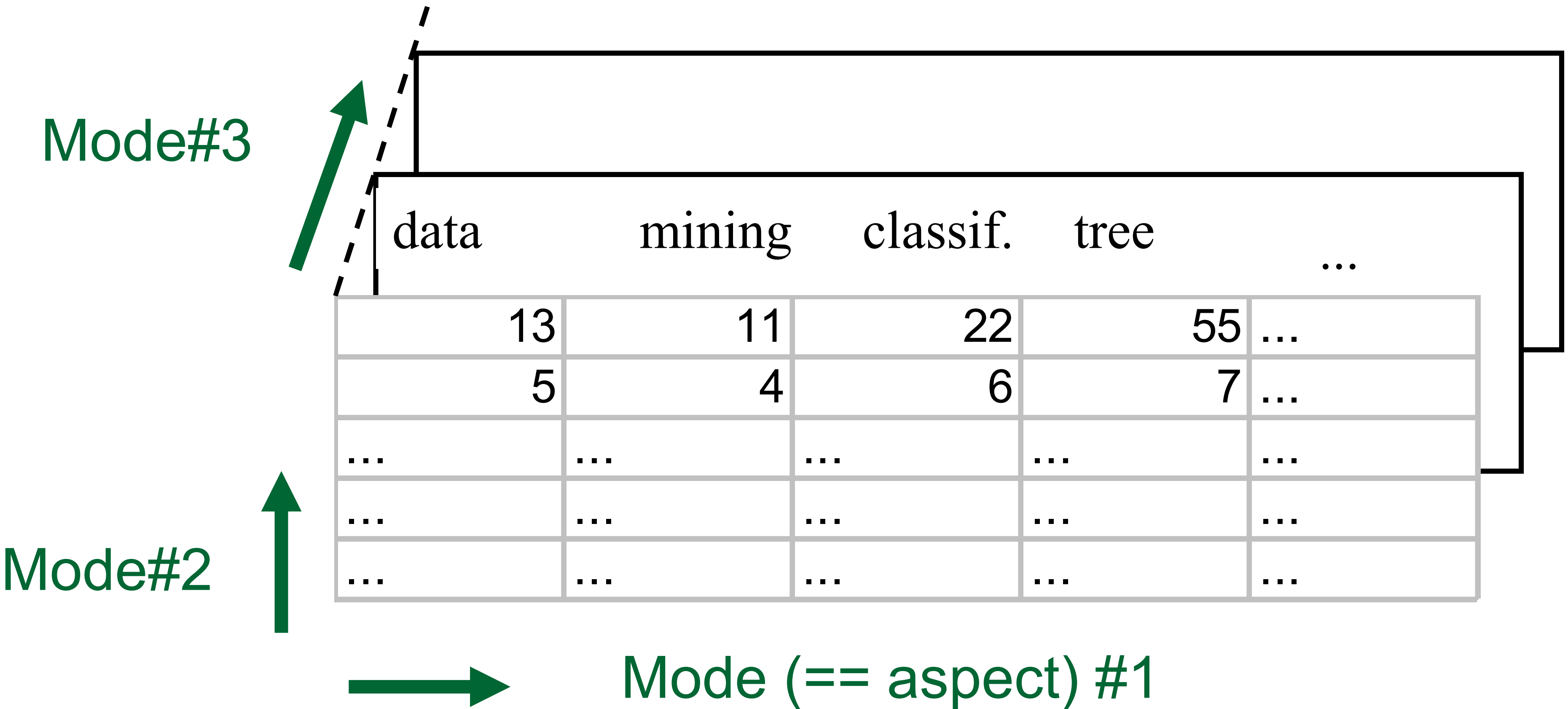
Motivation 2: Why tensor?

- A: N-D generalization of matrix:



Tensors are useful for 3 or more modes

Terminology: ‘mode’ (or ‘aspect’):

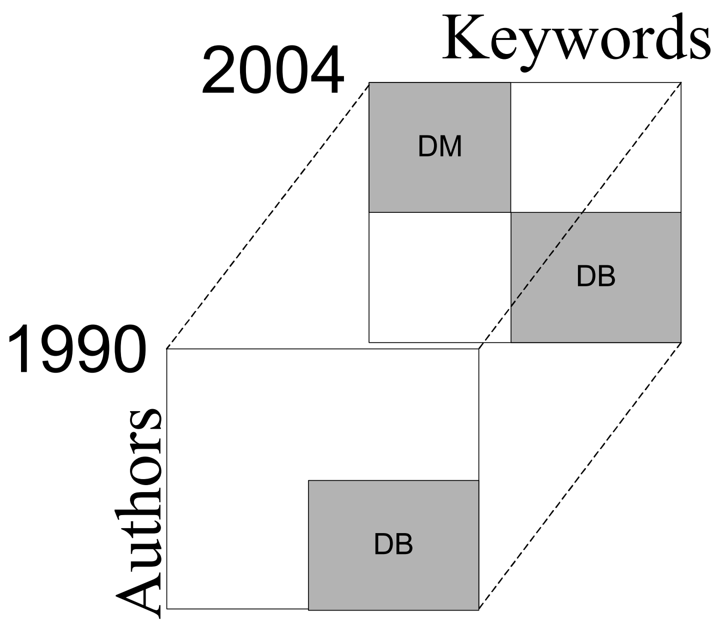


Motivating Applications

- Why matrices are important?
- Why tensors are useful?
  - P1: social networks
  - P2: web mining

P1: Social network analysis

- Previous research: people focus on static networks an d find community structures
- We plan to monitor the change of the community stru cture over time



P2: Web graph mining

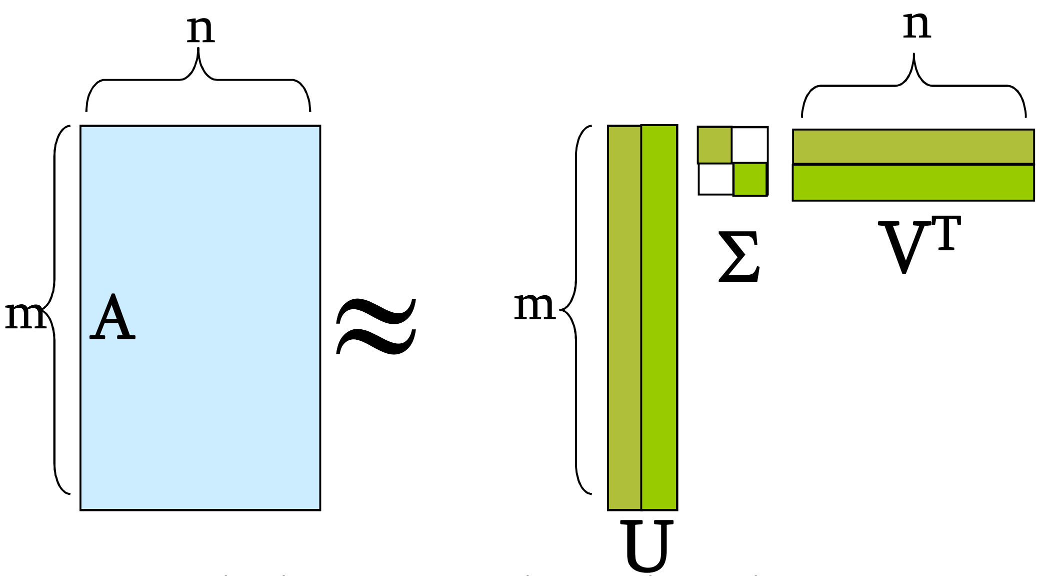
- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (TOPHITS)
    - context-sensitive hypergraph analysis

Outline

- ☑ Motivation - Definitions
- ☐ **Tensor tools**
  - Tensor basics
  - Parafac
  - Tucker
- ☐ Case Studies
- ☐ Conclusion

Reminder: SVD

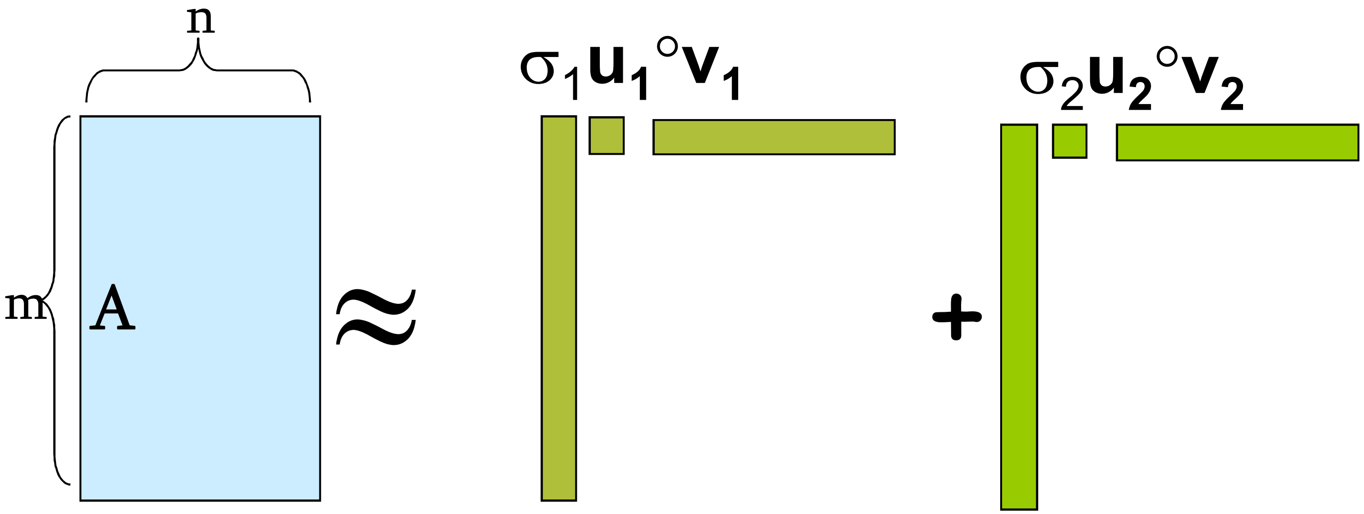
$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2

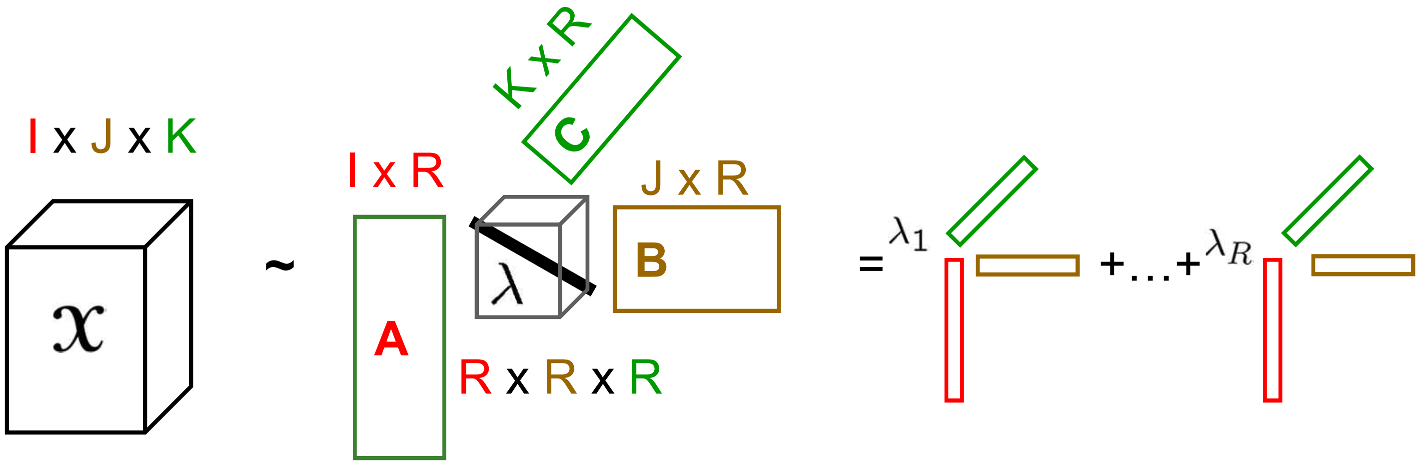
Reminder: SVD

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



- Best rank-k approximation in L2

Goal: extension to >=3 modes



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

Main points:

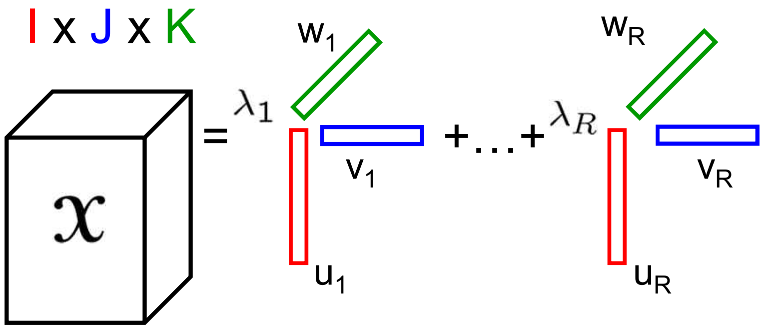
- 2 major types of tensor decompositions: PARAFAC and Tucker
- both can be solved with ``alternating least squares'' (ALS)
- Details follow

Specially Structured Tensors

- PARAFAC Tensor

$$\mathcal{X} = \sum_r \lambda_r \, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$$
$$\equiv [\![\lambda \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

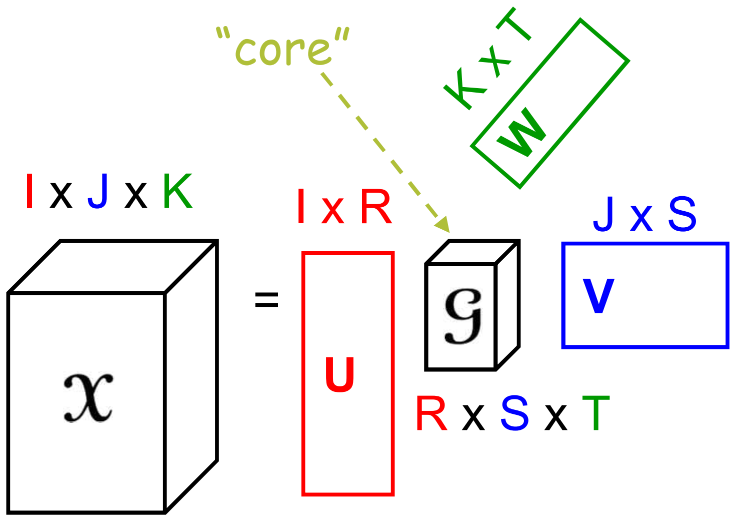
Our Notation



- Tucker Tensor

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}$$
$$= \sum_r \sum_s \sum_t g_{rst} \, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t$$
$$\equiv [\![\mathcal{G} \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

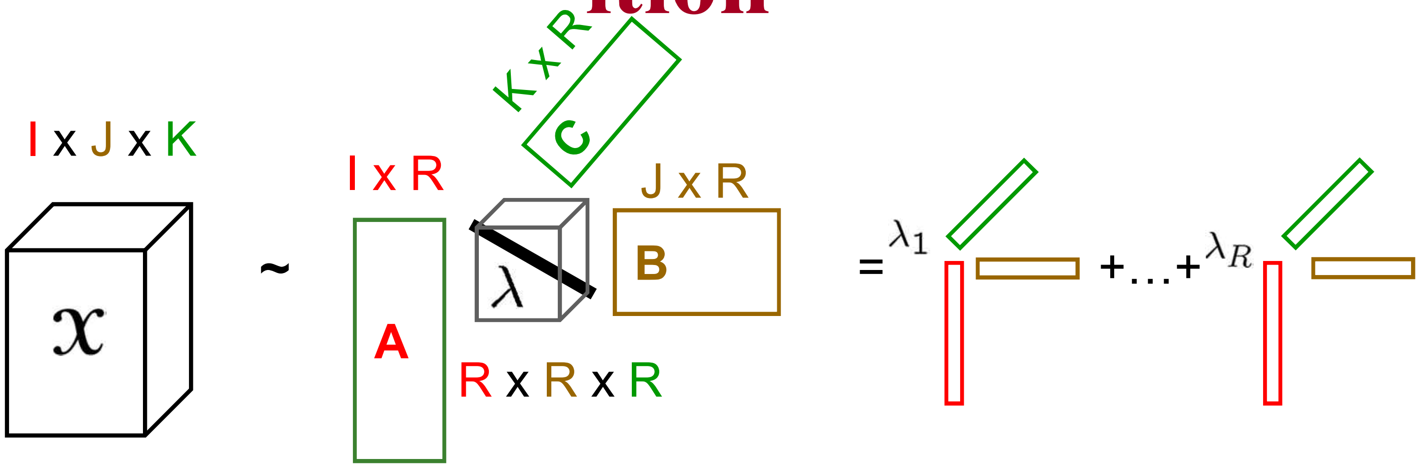
Our Notation



Outline

- ☑ Motivation - Definitions
- ☐ **Tensor tools**
  - Tensor basics
  - Parafac
  - Tucker
- ☐ Case Studies
- ☐ Conclusion

CANDECOMP/PARAFAC Decomposition



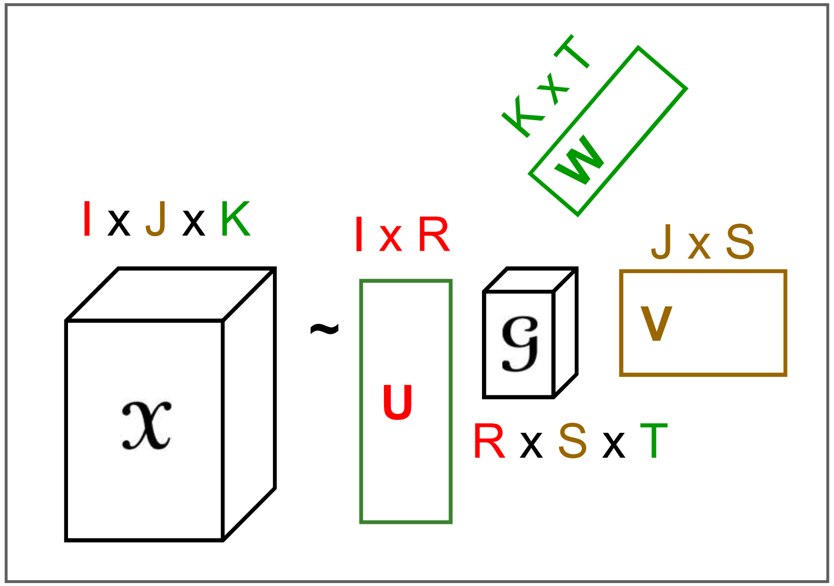
$$\mathcal{X} \approx [\![\lambda\,;\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_r \lambda_r\, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

- CANDECOMP = Canonical Decomposition (Carroll & Chang, 1970)
- PARAFAC = Parallel Factors (Harshman, 1970)
- Core is diagonal (specified by the vector λ)
- Columns of **A**, **B**, and **C** are not orthonormal
- If R is *minimum*, then R is called the **rank** of the tensor (Kruskal 1977)
- Can have rank($\mathcal{X}$) > min{I,J,K}

Outline

- ☑ Motivation - Definitions
- ☐ **Tensor tools**
  - Tensor basics
  - Parafac
  - Tucker
- ☐ Case Studies
- ☐ Conclusion

Tucker Decomposition - intuition



- author x keyword x conference
- U: author x author-group
- V: keyword x keyword-group
- W: conf. x conf-group
- $\mathcal{G}$: how groups relate to each other

Needs elaboration!

Intuition behind core tensor

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

$$
\begin{array}{cc}
 & n \\
m & \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}
\end{array}
$$

eg, terms x documents

$$
\begin{array}{cc}
 & k \\
m & \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix}
\end{array}
\begin{array}{cc}
 & l \\
k & \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}
\end{array}
\begin{array}{cc}
 & n \\
l & \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix}
\end{array}
=
\left[\begin{array}{ccc|ccc} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ \hline .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{array}\right]
$$

med. doc

cs doc

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

med. terms

cs terms

common terms

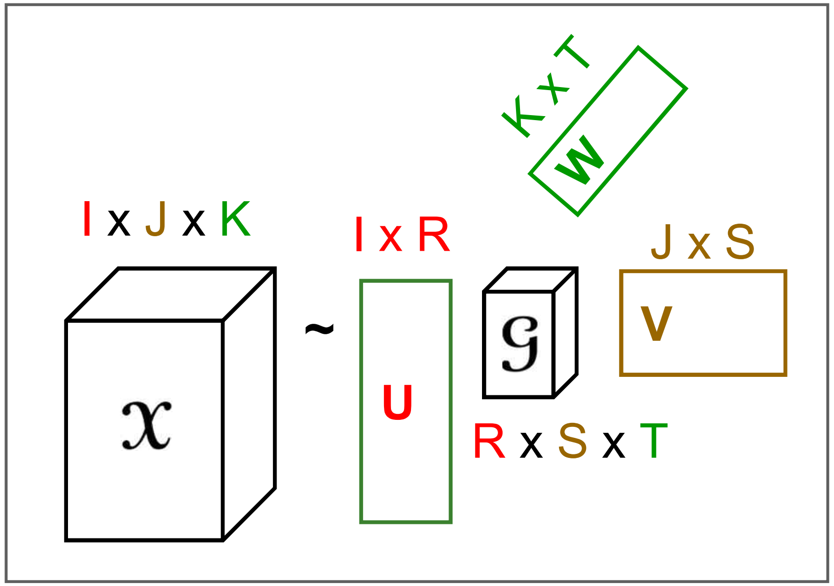
term group x doc. group

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

doc x doc group

term x term-group

Tucker Decomposition



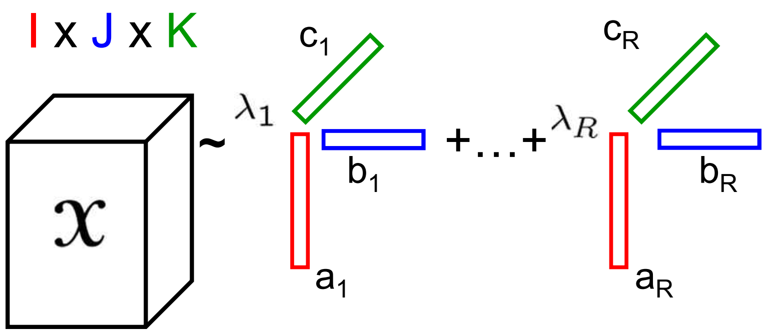
$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}$$
$$= \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t$$
$$\equiv [\![\mathcal{G} \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

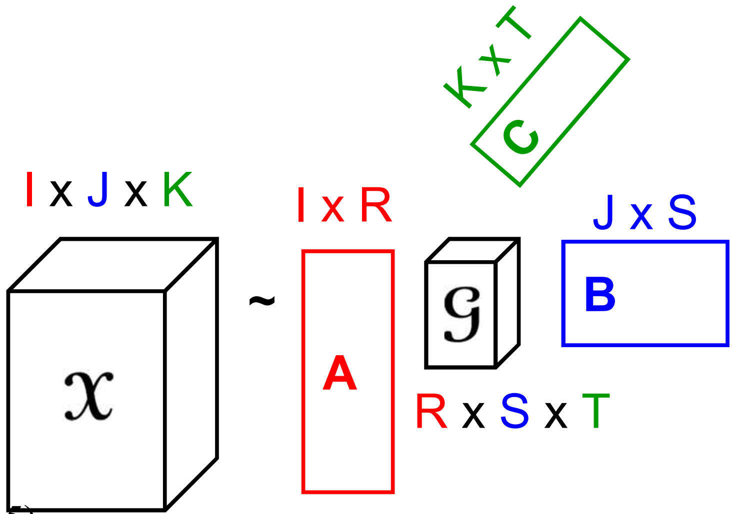
- Proposed by Tucker (1966)
- AKA: Three-mode factor analysis, three-mode PCA, orthogonal a rray decomposition
- **U**, **V**, and **W** generally assumed to be orthonormal (generally ass ume they have full column rank)
- $\mathcal{G}$ is <u>not</u> diagonal
- Not unique

IMPORTANT

PARAFAC vs. Tucker Decompositions

- PARAFAC
  - Sum of rank-1 components
  - No core, i.e., superdiagonal core
  - A, B, C may have linearly dependent columns
  - Generally unique
- Tucker
  - Many interactions from groups
  - Core G may be dense
  - A, B, C generally orthonormal
  - Not unique





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, gradient desc ent, …

- Toolbox: from Tamara Kolda:

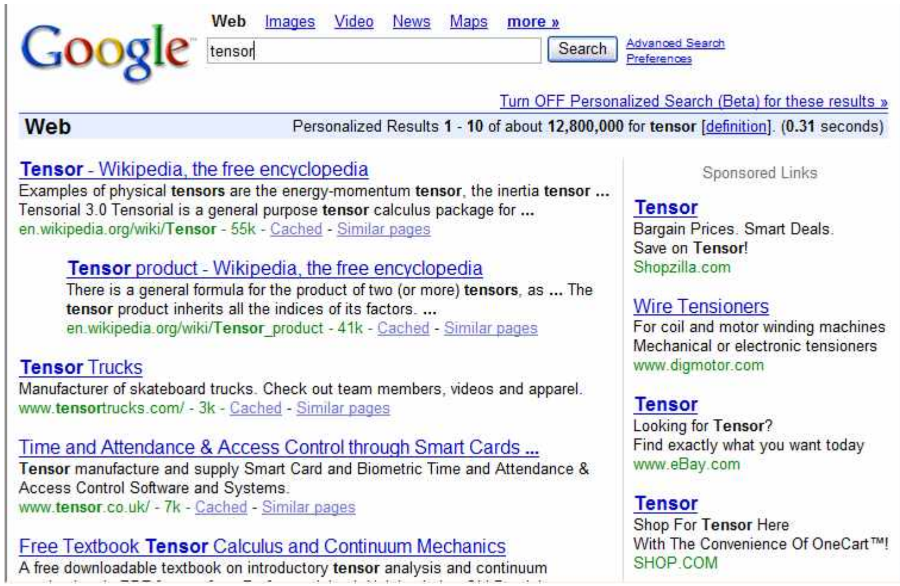
http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/

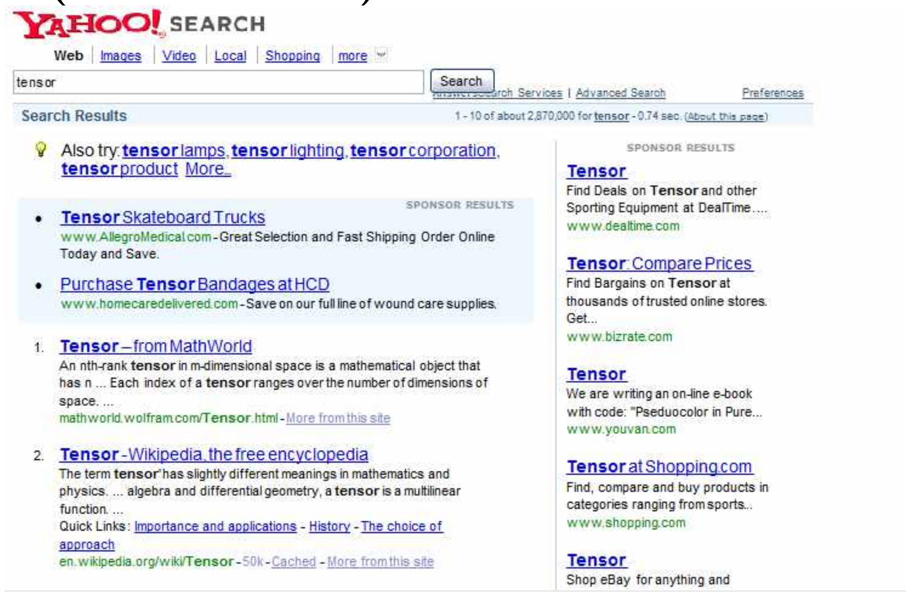
Outline

- ☑ Motivation - Definitions
- ☑ Tensor tools
- ☐ **Case Studies**
  - HITS
  - GigaTensor
- ☐ Conclusion

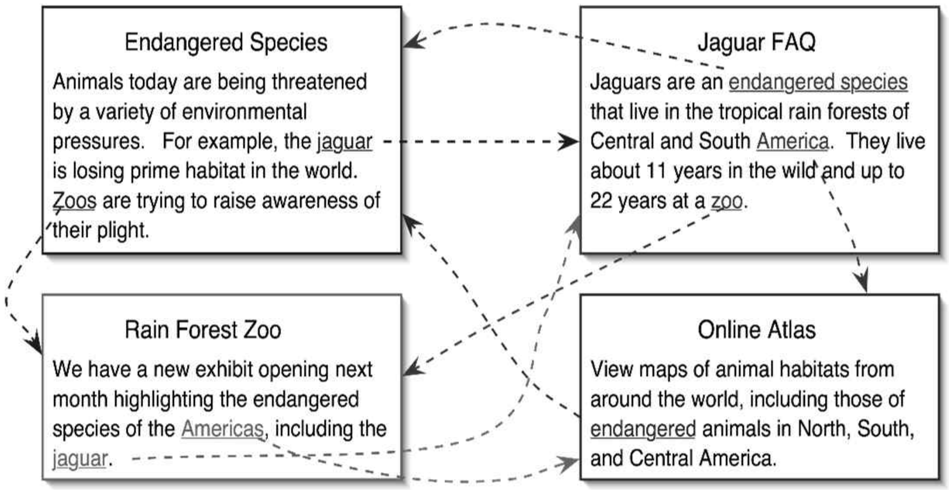
P1: Web graph mining

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





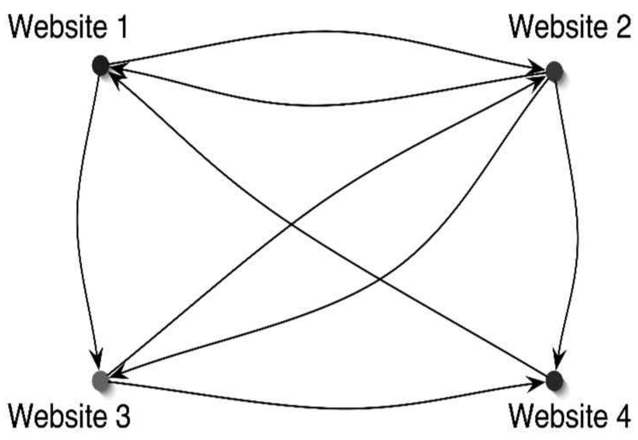
Kleinberg's Hubs and Authorities (the HITS method)

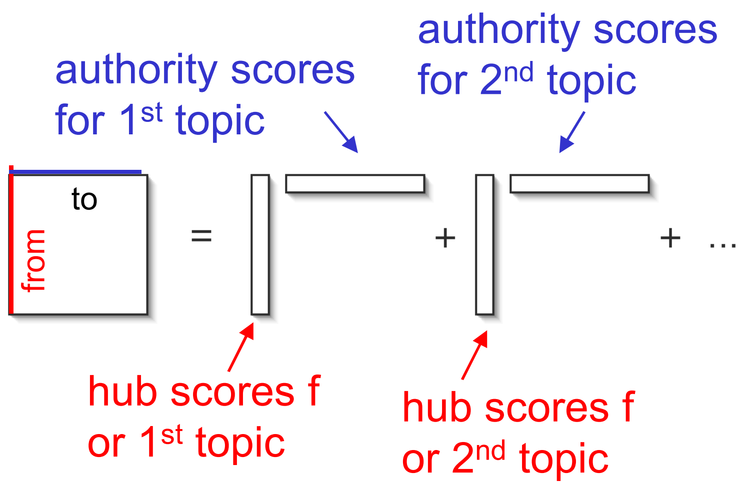


Sparse adjacency matrix and its SVD:

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

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





HITS Authorities on Sample Data

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

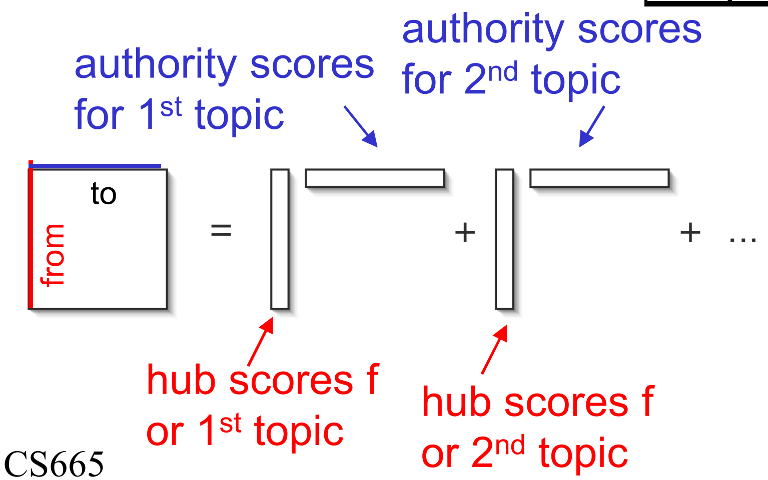
2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh
.06	www.lehigha
.06	www.lehighs
.02	www.bethleh
.02	www.adobe.c
.02	lewisweb.cc.
.02	www.leo.lehi
.02	www.distanc
.02	fp1.cc.lehigh

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

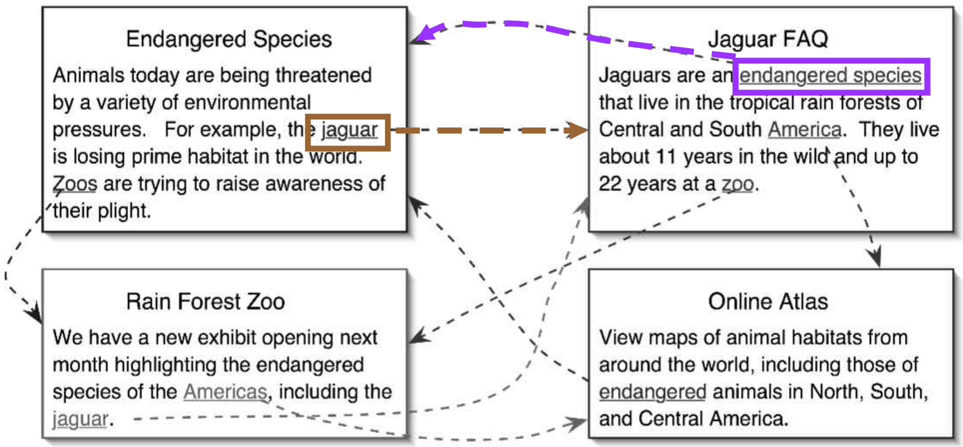
4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state
.22	www.gsa.g
.20	www.ssa.g
.16	www.censu
.14	www.govbe
.13	www.kids.g
.13	www.usdoj

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

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

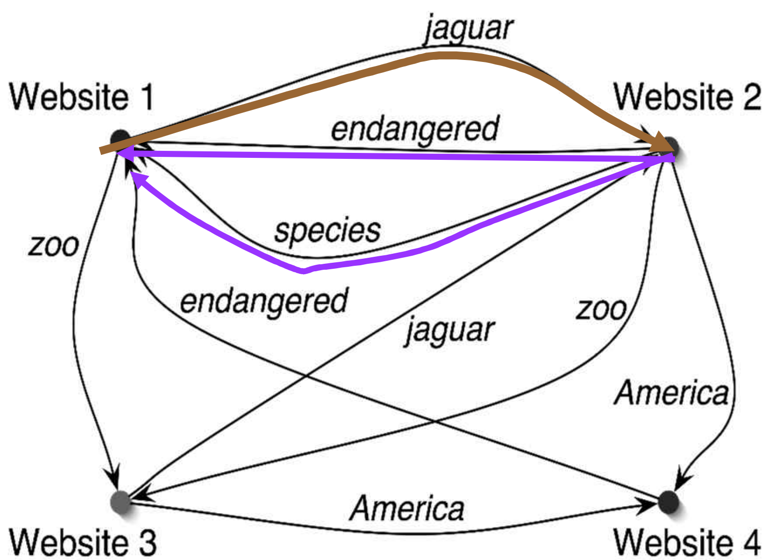


Three-Dimensional View of the Web

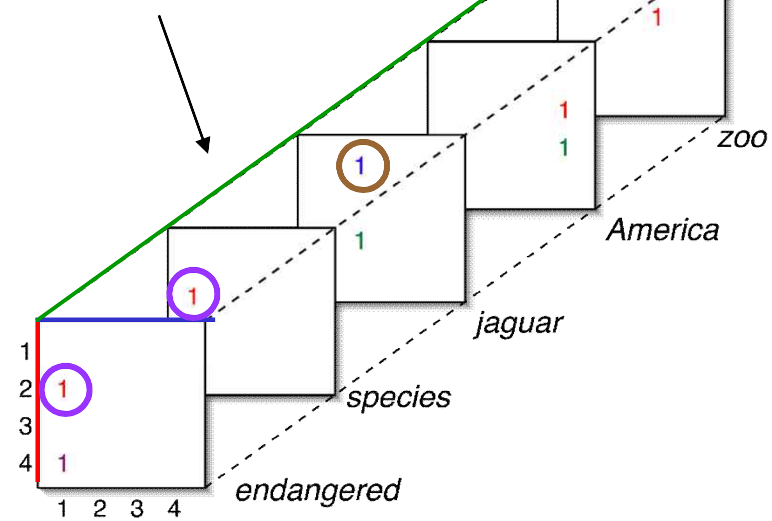


[Kolda, Bader, Kenny, ICDM05]

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



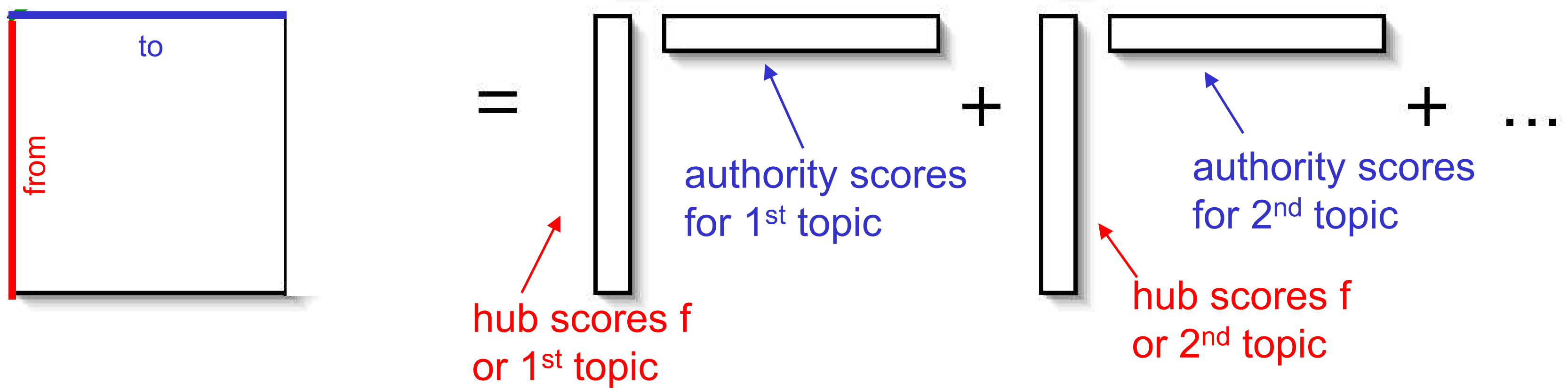
Observe that this tensor is very sparse!



Topical HITS (TOPHITS)

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

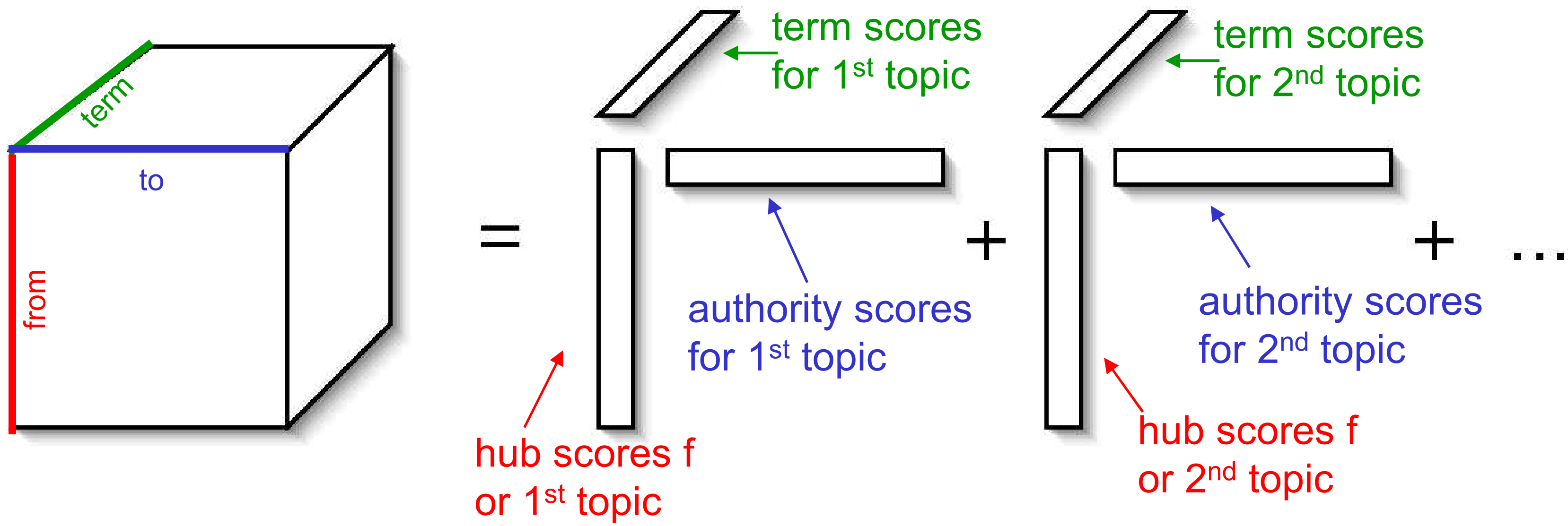
$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



Topical HITS (TOPHITS)

<u>Main Idea</u>: Extend the idea behind the HITS model to incorporate t erm (i.e., topical) information.

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



TOPHITS Terms & Authorities on Sample Data

TERM **AUTHORITY**

1st Principal Factor			
.23	JAVA	.86	java.sun.com
.18	SUN	.38	developers.sun.com
.17	PLATF		
.16	SOLAR		
.16	DEVEL		
.15	EDITIO		
.15	DOWN		
.14	INFO		
.12	SOFTW		
.12	NO-RE		

2nd Principal Factor			
.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACUL		
.16	SEARC		
.16	NEWS		
.16	LIBRAR		
.16	COMP		
.12	LEHIG		

3rd Principal Factor			
.15	NO-READABLE-TEXT	.97	www.ibm.com
.15	IBM	.18	www.alphaworks.ibm.com
.12	SERVI		
.12	WEBS		
.12	WEB		
.11	DEVEL		
.11	LINUX		
.11	RESOU		
.11	TECHN		
.10	DOWN		

4th Principal Factor			
.26	INFORMATION	.87	www.pueblo.gsa.gov
.24	FEDERAL	.24	www.irs.gov
.23	CITIZE		
.22	OTHER		
.19	CENTE		
.19	LANGU		
.15	U.S		
.15	PUBLI		
.14	CONSU		
.13	FREE		

6th Principal Factor			
.26	PRESIDENT	.87	www.whitehouse.gov
.25	NO-READABLE-TEXT	.18	www.irs.gov
.25	BUSH		
.25	WELC		
.17	WHITE		
.16	U.S		
.15	HOUSE		
.13	BUDG		
.13	PRESI		
.11	OFFIC		

12th Principal Factor			
.75	OPTIMIZATION	.35	www.palisade.com
.58	SOFTWARE	.35	www.solver.com
.08	DECIS		
.07	NEOS		
.06	TREE		
.05	GUIDE		
.05	SEARC		
.05	ENGIN		
.05	CONTR		
.05	ILOG		

13th Principal Factor			
.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACRO		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

16th Principal Factor			
.50	WEATHER	.81	www.weather.gov
.24	OFFICE	.41	www.spc.noaa.gov
.23	CENTER	.30	lwf.ncdc.noaa.gov
.19	NO-RE		
.17	ORGA		
.15	NWS		
.15	SEVER		
.15	FIRE		
.15	POLIC		
.14	CLIMA		

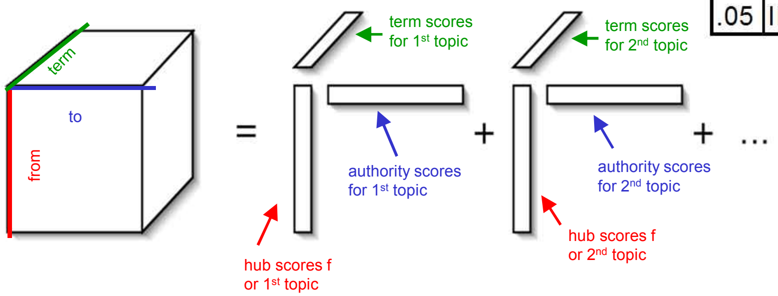
19th Principal Factor			
.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

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

$$\mathbf{A}_{ijk} = \begin{cases} 1 + \log(\mathbf{C}_{ijk}) & \text{if } \mathbf{C}_{ijk} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

C_{ijk}: # of links from i to j with term k

Tensor PARAFAC



Outline

- ☑ Motivation - Definitions
- ☑ Tensor tools
- ➡ ☐ **Case Studies**
  - HITS
  - ➡ GigaTensor
- ☐ Conclusion

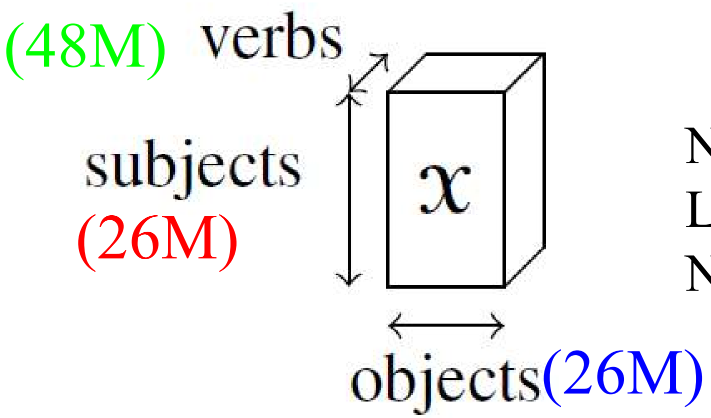
U Kang, Evangelos Papalexakis, Abhay Harpale, and Christos Faloutsos. GigaTensor: Scaling Tensor Analysis Up By 100 Times - Algorithms and Discoveries, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2012, Beijing, China.

P2: N.E.L.L. analysis

- NELL: Never Ending Language Learner
  - Q1: dominant concepts / topics?
  - Q2: synonyms for a given new phrase?

"Eric Clapton plays guitar"

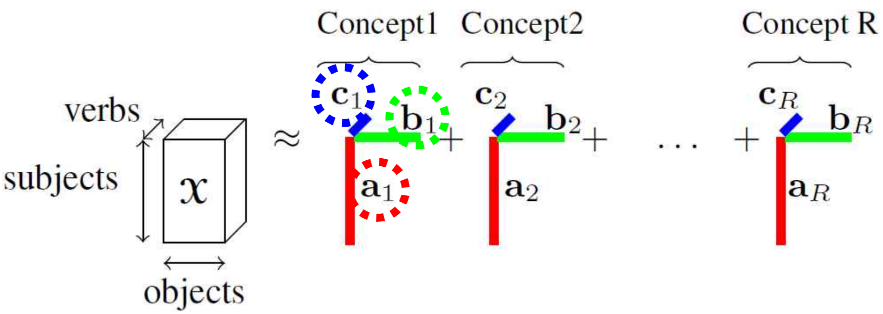
"Barrack Obama is the president of U.S."



NELL (Never Ending Language Learner)
Nonzeros =144M

A1: Concept Discovery

- Concept Discovery in Knowledge Base



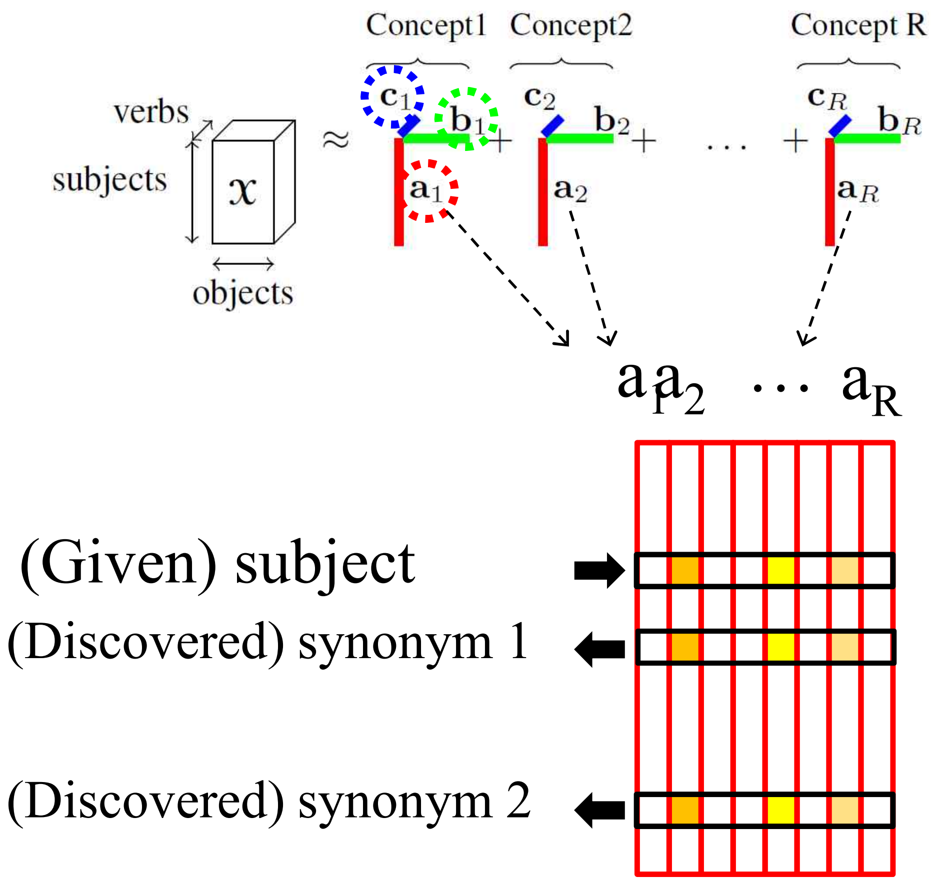
Noun Phrase 1	Noun Phrase 2	Context
Concept 1: "Web Protocol"		
internet	protocol	'np1' 'stream' 'np2'
file	software	'np1' 'marketing' 'np2'
data	suite	'np1' 'dating' 'np2'
Concept 2: "Credit Cards"		
credit	information	'np1' 'card' 'np2'
Credit	debt	'np1' 'report' 'np2'
library	number	'np1' 'cards' 'np2'
Concept 3: "Health System"		
health	provider	'np1' 'care' 'np2'
child	providers	'np' 'insurance' 'np2'
home	system	'np1' 'service' 'np2'
Concept 4: "Family Life"		
life	rest	'np2' 'of' 'my' 'np1'
family	part	'np2' 'of' 'his' 'np1"
body	years	'np2' 'of' 'her' 'np1'

A1: Concept Discovery

Noun Phrase 1	Noun Phrase 2	Context
Concept 1: "Web Protocol"		
internet	protocol	‘np1’ ‘stream’ ‘np2’
file	software	‘np1’ ‘marketing’ ‘np2’
data	suite	‘np1’ ‘dating’ ‘np2’
Concept 2: "Credit Cards"		
credit	information	‘np1’ ‘card’ ‘np2’
Credit	debt	‘np1’ ‘report’ ‘np2’
library	number	‘np1’ ‘cards’ ‘np2’
Concept 3: "Health System"		
health	provider	‘np1’ ‘care’ ‘np2’
child	providers	’np’ ‘insurance’ ‘np2’
home	system	‘np1’ ‘service’ ‘np2’
Concept 4: "Family Life"		
life	rest	‘np2’ ‘of’ ‘my’ ‘np1’
family	part	‘np2’ ‘of’ ‘his’ ‘np1”
body	years	‘np2’ ‘of’ ‘her’ ‘np1’

A2: Synonym Discovery

- Synonym Discovery in Knowledge Base



(Given) **Noun Phrase**	(Discovered) **Potential Synonyms**
pollutants	dioxin, sulfur dioxide, greenhouse gases, particulates, nitrogen oxide, air pollutants, cholesterol
disabilities	infections, dizziness, injuries, diseases, drowsiness, stiffness, injuries
vodafone	verizon, comcast
Christian history	European history, American history, Islamic history, history
disbelief	dismay, disgust, astonishment
cyberpunk	online-gaming
soul	body

A2: Synonym Discovery

(Given) Noun Phrase	(Discovered) Potential Synonyms
pollutants	dioxin, sulfur dioxide, greenhouse gases, particulates, nitrogen oxide, air pollutants, cholesterol
disabilities	infections, dizziness, injuries, diseases, drowsiness, stiffness, injuries
vodafone	verizon, comcast
Christian history	European history, American history, Islamic history, history
disbelief	dismay, disgust, astonishment
cyberpunk	online-gaming
soul	body

Outline

- ☑ Motivation - Definitions
- ☑ Tensor tools
- ☑ Case Studies
- ☐ **Conclusion**

Conclusions

- Real data are often in high dimensions with multi ple aspects (modes)
- Matrices and tensors provide elegant theory and al gorithms