Carnegie Mellon

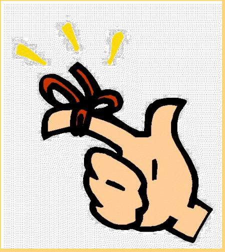
15-826: Multimedia (Databases) and Data Mining

Lecture #19: Tensor decompositions

C. Faloutsos

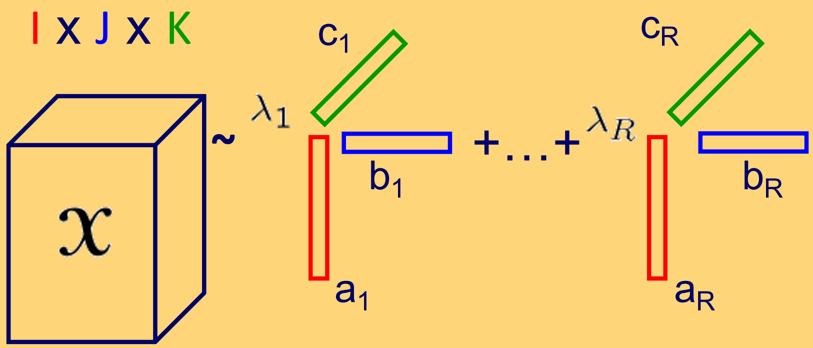
Problem

- Q: who-calls-whom-when – patterns?
  - Triplets (source-ip, dest-ip, port#)
  - KB (subject, verb, object)



Conclusions

- Q: who-calls-whom-when – patterns?
  – Triplets (source-ip, dest-ip, port#)
  – KB (subject, verb, object)
- A: Tensor analysis (PARAFAC)
  – http://www.tensortoolbox.org/



Must-read Material

- [Graph-Textbook] Ch.16.

- Tensors survey: Papalexakis, Faloutsos, Sidiropoulos Tensor for Data Mining and Data Fusion: Models, Applications, and Scalable Algorithms ACM Trans. on Intelligent Systems and Technology, 8,2, Oct. 2016. (local copy)

Outline

Goal: ‘Find similar / interesting things’

- Intro to DB
- Indexing - similarity search
- Data Mining

Carnegie Mellon

Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- text
- Singular Value Decomposition (SVD)
  - …
  - Tensors
- multimedia
- ...

Outline

- Motivation - Definitions
- Tensor tools
- Case studies

Carnegie Mellon

Most of foils by

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

- Prof. Jimeng Sun (UIUC)
- https://cs.illinois.edu/about/people/faculty/jimeng

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

Motivation 1: 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' 17	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:

KDD' 15

KDD' 16

KDD' 17

	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'):

Mode#3

data	mining	classif.	tree	...
13	11	22	55	...
5	4	6	7	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Mode#2

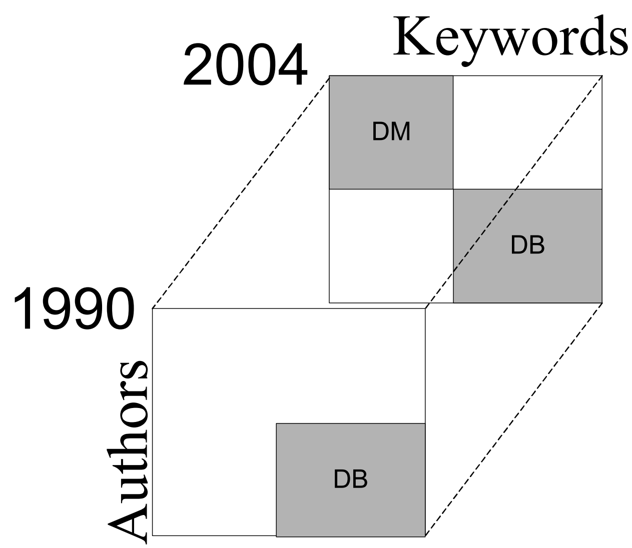
Mode (== aspect) #1

Motivating Applications

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

P1: Social network analysis

- Traditionally, people focus on static networks and find community structures
- We plan to monitor the change of the community structure 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

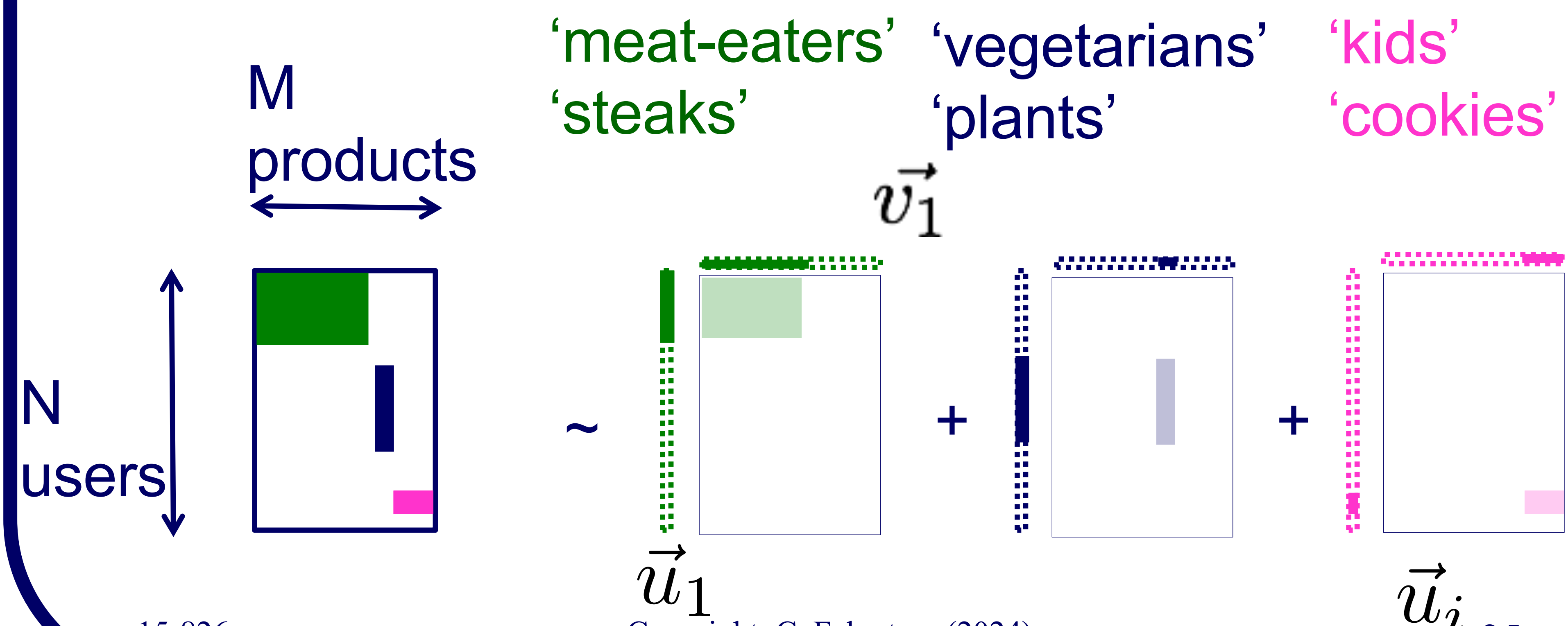
- Motivation – Definitions
- Tensor tools
- Case studies

- Tensor Basics
- Tucker
- PARAFAC

Tensor Basics

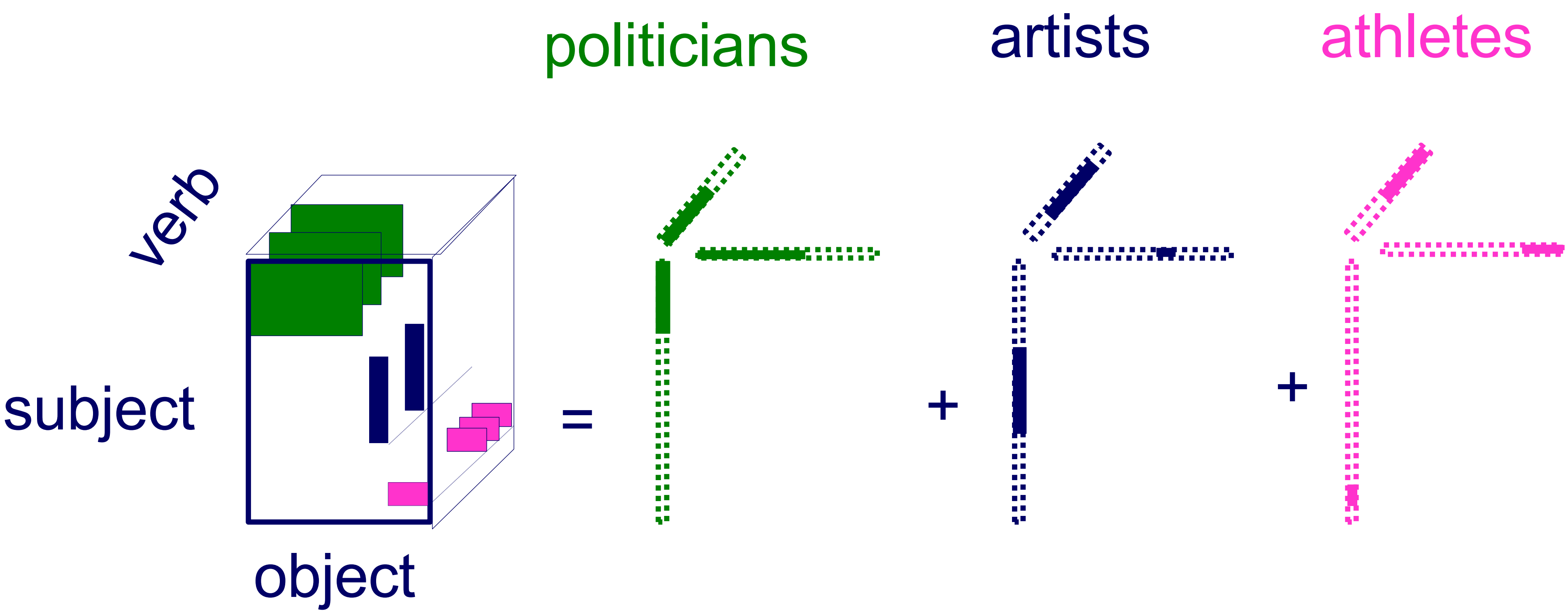
Answer to both: tensor factorization

- Recall: (SVD) matrix factorization: finds blocks



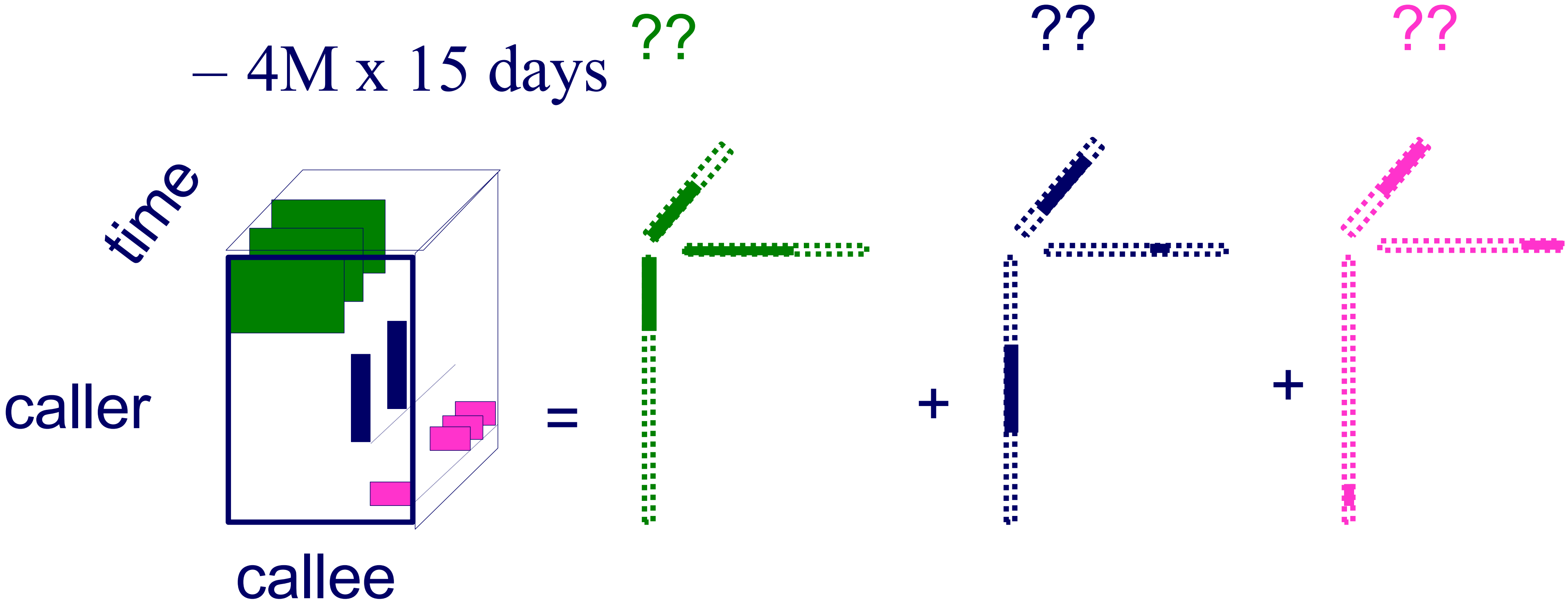
Answer to both: tensor factorization

- PARAFAC decomposition



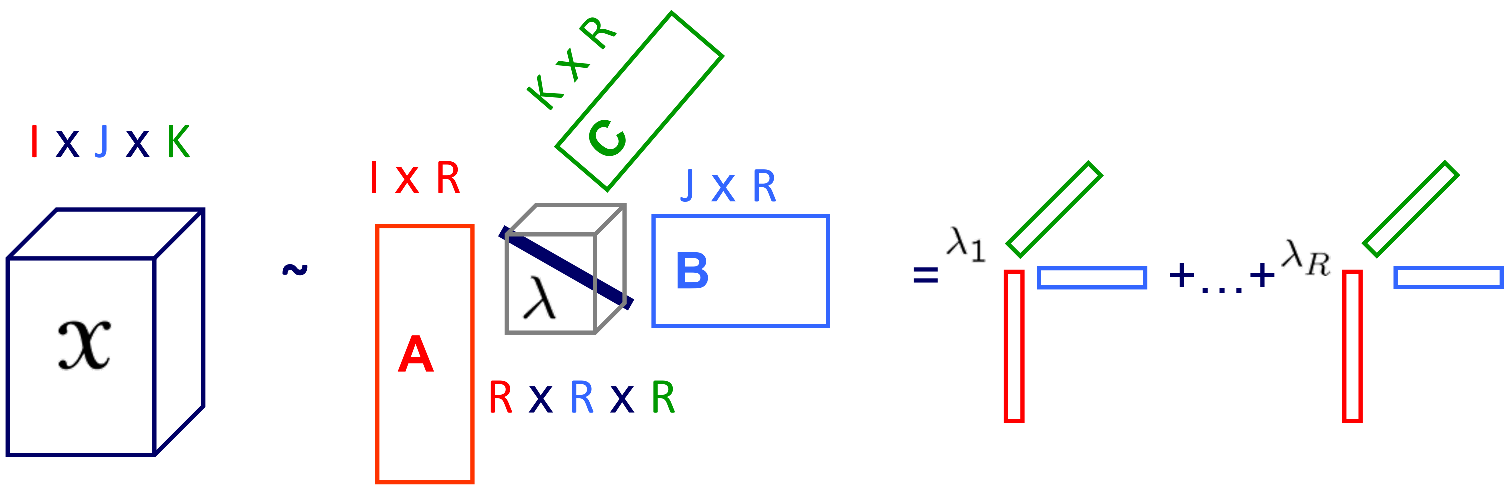
Answer: tensor factorization

- PARAFAC decomposition
- Results for who-calls-whom-when
  - 4M x 15 days



Carnegie Mellon

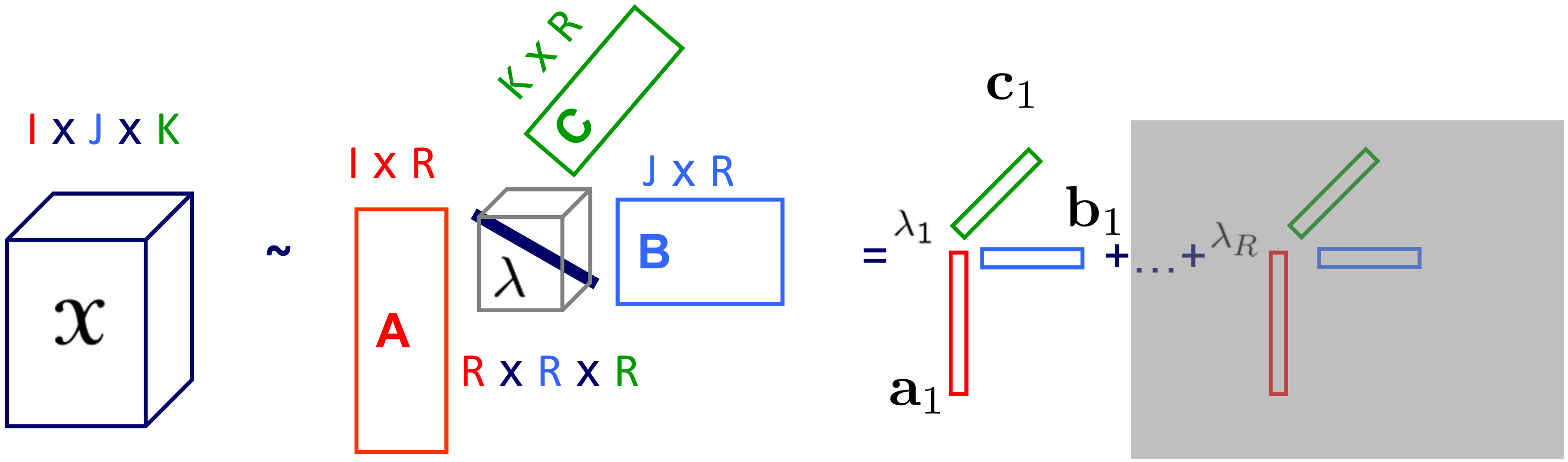
Goal: extension to >=3 modes



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

Example of outer product ‘o’:

Goal: extension to >=3 modes

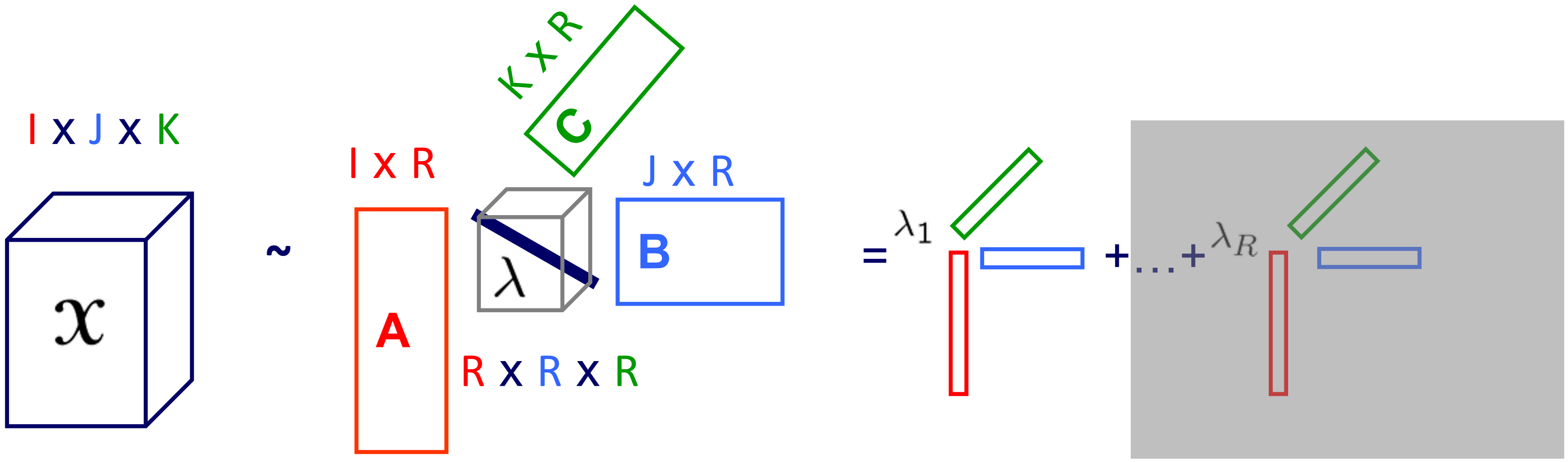


$$\mathcal{X} \approx [\![\lambda\, ; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_r \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Suppose
R=1
a1=(1,2,3,4)
b1=(2,2,2)
c1=(10,11)
λ1=7

X(1,1,1)=?
X(3,1,2)=?
X(5,1,1)=?

Goal: extension to >=3 modes

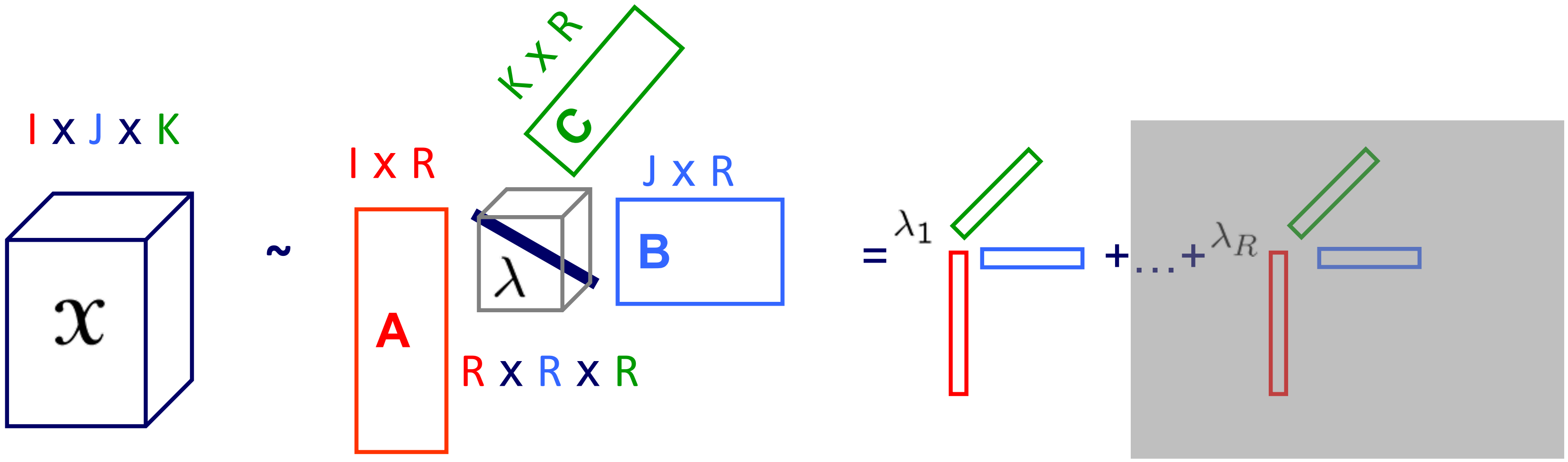


$$\mathcal{X} \approx [\![\lambda\ ; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_r \lambda_r\ \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Suppose
r=1
a1=(**1**,2,3,4)
b1=(**2**,2,2)
c1=(**10**,11)
λ1=**7**

X(1,1,1)=7 *1*2*10
X(3,1,2)=?
X(5,1,1)=?

Goal: extension to >=3 modes



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

Suppose

r=1

a1=(1,2,**3**,4)

b1=(**2**,2,2)

c1=(10,**11**)

λ1=**7**

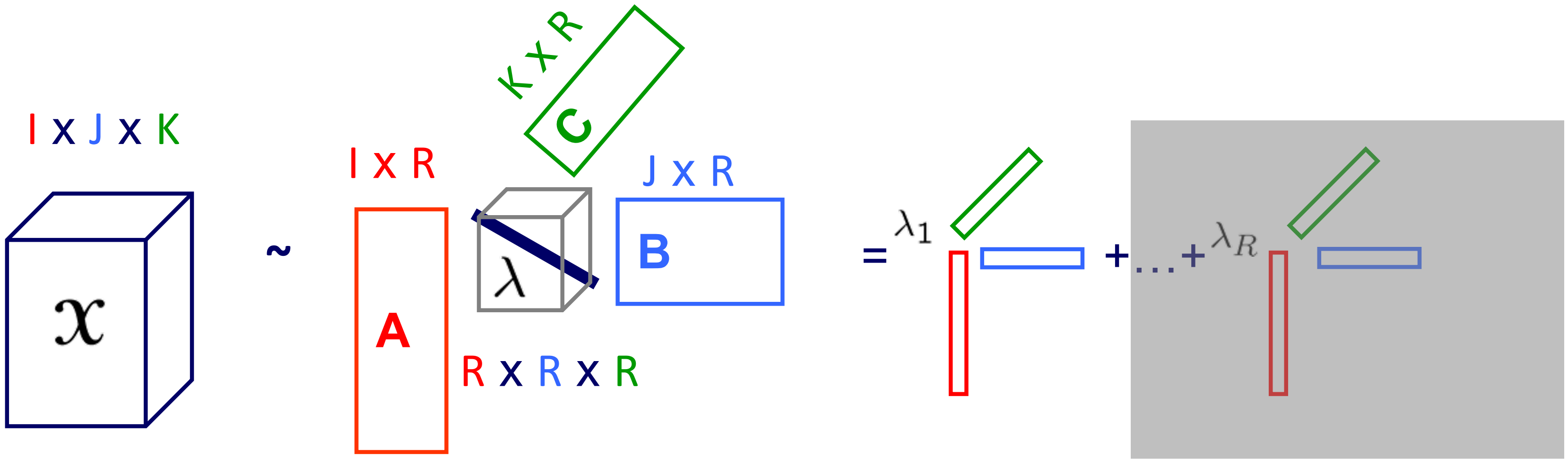
X(1,1,1)=7*1*2*10

X(3,1,2)= 7*3*2*11

X(5,1,1)= ??

Carnegie Mellon

Goal: extension to >=3 modes



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

Suppose
r=1
a1=(1,2,**3**,4)
b1=(**2**,2,2)
c1=(10,**11**)
λ1=**7**

X(1,1,1)=7*1*2*10
X(3,1,2)= 7*3*2*11
X(5,1,1)= N/A - TRICK QUESTION

Main points:

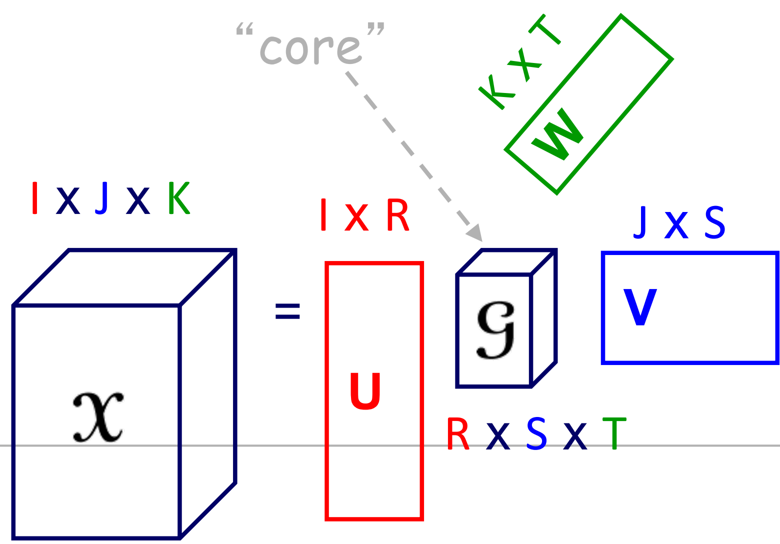
- 2 major types of tensor decompositions: PARAFAC and Tucker
- both can be solved with ``alternating least squares'' (ALS)
- Details follow

Specially Structured Tensors

Specially Structured Tensors

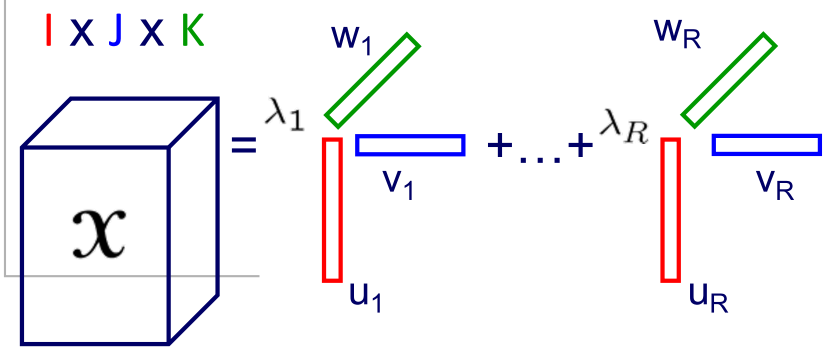
- Tucker Tensor

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



- Kruskal Tensor

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



details

Specially Structured Tensors

- Tucker Tensor

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

In matrix form:

$$\mathbf{X}_{(1)} = \mathbf{U}\mathbf{G}_{(1)}(\mathbf{W} \otimes \mathbf{V})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{V}\mathbf{G}_{(2)}(\mathbf{W} \otimes \mathbf{U})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{W}\mathbf{G}_{(3)}(\mathbf{V} \otimes \mathbf{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathbf{W} \otimes \mathbf{V} \otimes \mathbf{U})\mathrm{vec}(\mathcal{G})$$

- Kruskal Tensor

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

In matrix form:

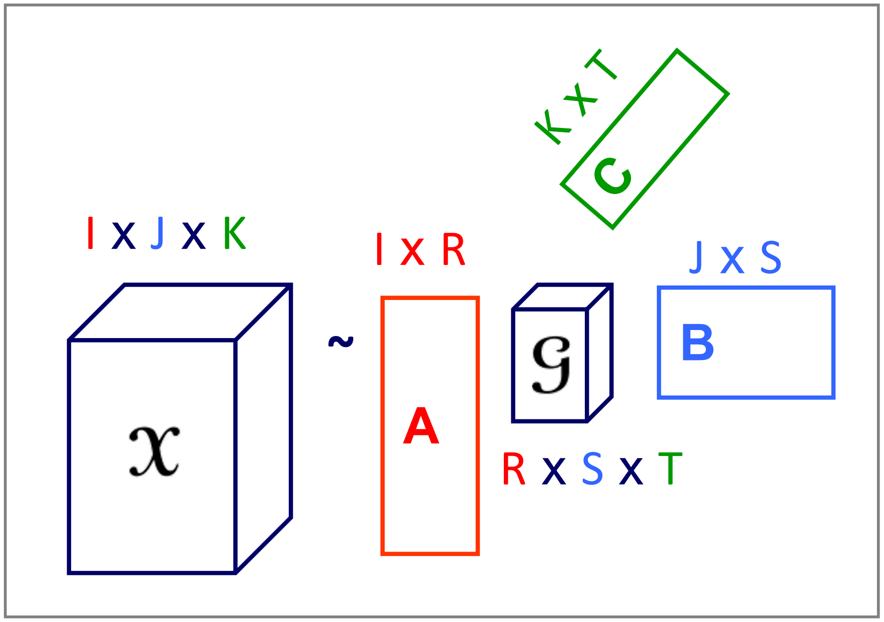
Let $\mathbf{\Lambda} = \mathrm{diag}(\lambda)$

$$\mathbf{X}_{(1)} = \mathbf{U} \, \mathbf{\Lambda} \, (\mathbf{W} \odot \mathbf{V})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{V} \, \mathbf{\Lambda} \, (\mathbf{W} \odot \mathbf{U})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{W} \, \mathbf{\Lambda} \, (\mathbf{V} \odot \mathbf{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathbf{W} \odot \mathbf{V} \odot \mathbf{U}) \, \lambda$$

Tensor Decompositions

Tucker Decomposition - intuition



- author x keyword x conference
- A: author x author-group
- B: keyword x keyword-group
- C: 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

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

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

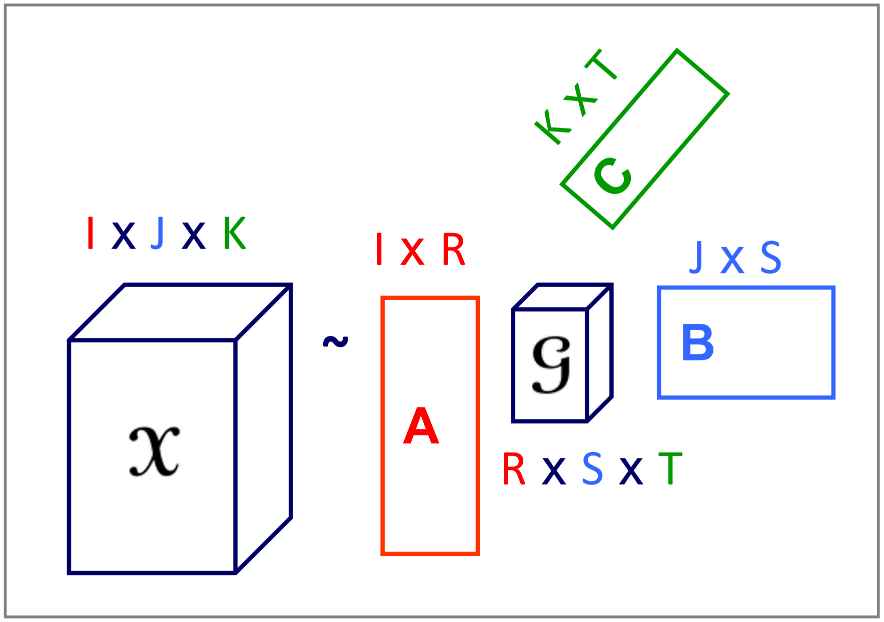
term group x doc. group

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

doc x doc group

term x term-group

Tucker Decomposition



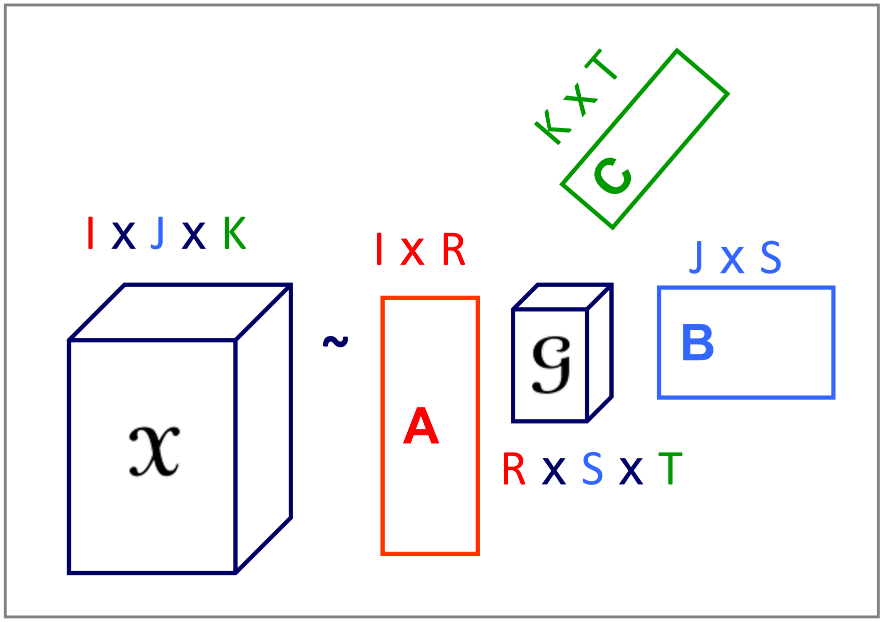
$$\mathcal{X} \approx [\![\mathcal{G} \,; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

- Proposed by Tucker (1966)
- AKA: Three-mode factor analysis, three-mode PCA, orthogonal array decomposition
- **A**, **B**, and **C** generally assumed to be orthonormal (generally assume they have full column rank)
- $\mathcal{G}$ is <u>not</u> diagonal
- Not unique

Carnegie Mellon

details

Tucker Decomposition



$$\mathcal{X} \approx [\![\mathcal{G} \,; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Given A, B, C, the optimal core is:

$$\mathcal{G} = [\![\mathcal{X} \,; \mathbf{A}^{\dagger}, \mathbf{B}^{\dagger}, \mathbf{C}^{\dagger}]\!]$$

- Proposed by Tucker (1966)
- AKA: Three-mode factor analysis, three-mode PCA, orthogonal array decomposition
- **A**, **B**, and **C** generally assumed to be orthonormal (generally assume they have full column rank)
- $\mathcal{G}$ is not diagonal
- Not unique

Recall the equations for converting a tensor to a matrix

$$\mathbf{X}_{(1)} = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathrm{vec}(\mathcal{X}) = (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathcal{G})$$

details

Kronecker product

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 10 & 20 & 30 \end{bmatrix}$$

m1 x n1

m2 x n2

*m1*m2 x n1*n2*

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} 1 * \mathbf{B} & 2 * \mathbf{B} \\ 3 * \mathbf{B} & 4 * \mathbf{B} \end{bmatrix}$$

$$= \begin{bmatrix} 1*10 & 1*20 & 1*30 & 2*10 & 2*20 & 2*30 \\ 3*10 & 3*20 & 3*30 & 4*10 & 4*20 & 4*30 \end{bmatrix}$$

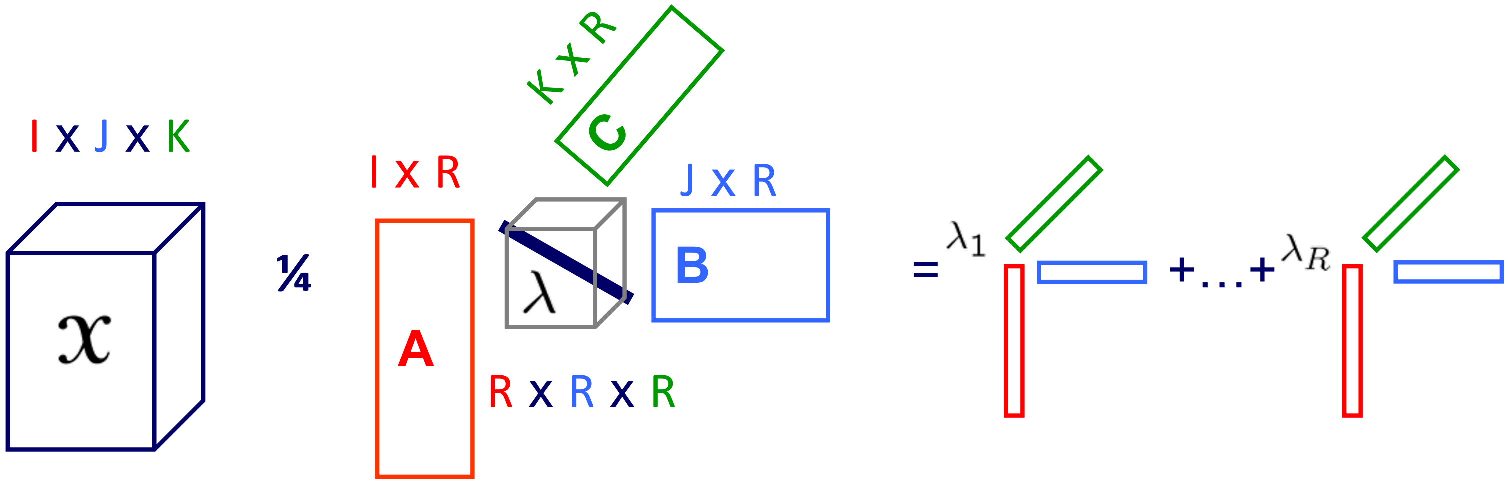
Outline

- Motivation – Definitions
- Tensor tools
- Case studies

- Tensor Basics
- Tucker
- PARAFAC

Carnegie Mellon

CANDECOMP/PARAFAC Decomposition



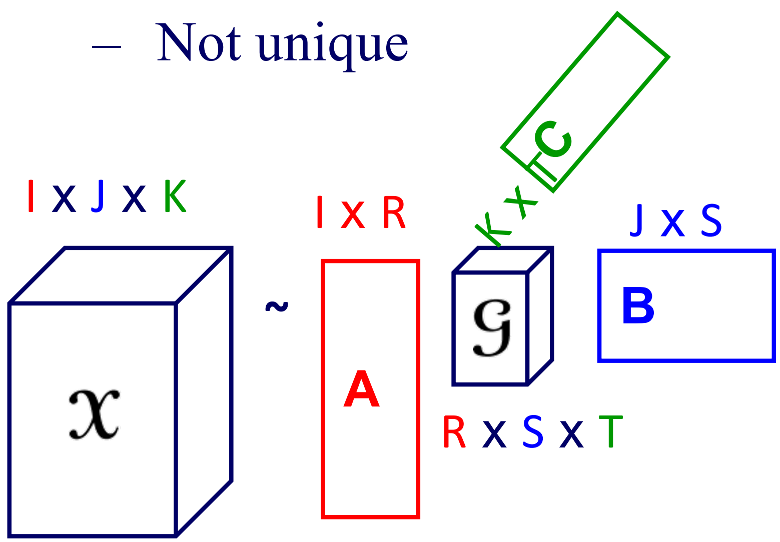
$$\mathcal{X} \approx [\![\lambda \,;\, \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_r \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

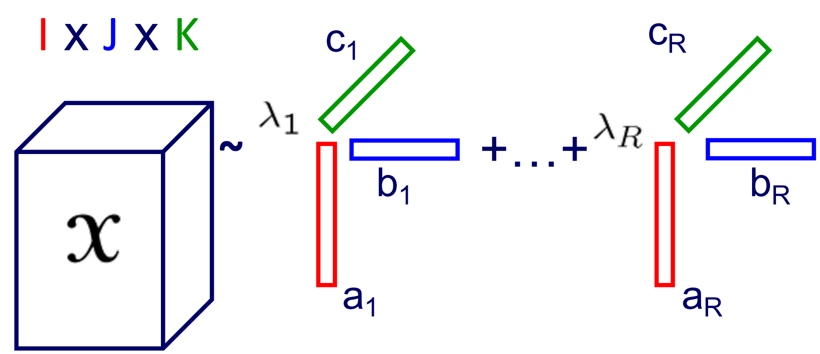
- CANDECOMP = Canonical Decomposition (Carroll & Chang, 1970)
- PARAFAC = Parallel Factors (Harshman, 1970)
- Core is <u>diagonal</u> (specified by the vector λ)
- Columns of **A**, **B**, and **C** are <u>not</u> orthonormal
- If R is <u>*minimal*</u>, then R is called the **rank** of the tensor (Kruskal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

IMPORTANT

Tucker vs. PARAFAC Decompositions

- Tucker
  - Variable transformation in each mode
  - Core G may be dense
  - A, B, C generally orthonormal
  - Not unique
- PARAFAC
  - Sum of rank-1 components
  - No core, i.e., superdiagonal core
  - A, B, C may have linearly dependent columns
  - Generally 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

- Toolbox: from Tamara Kolda:

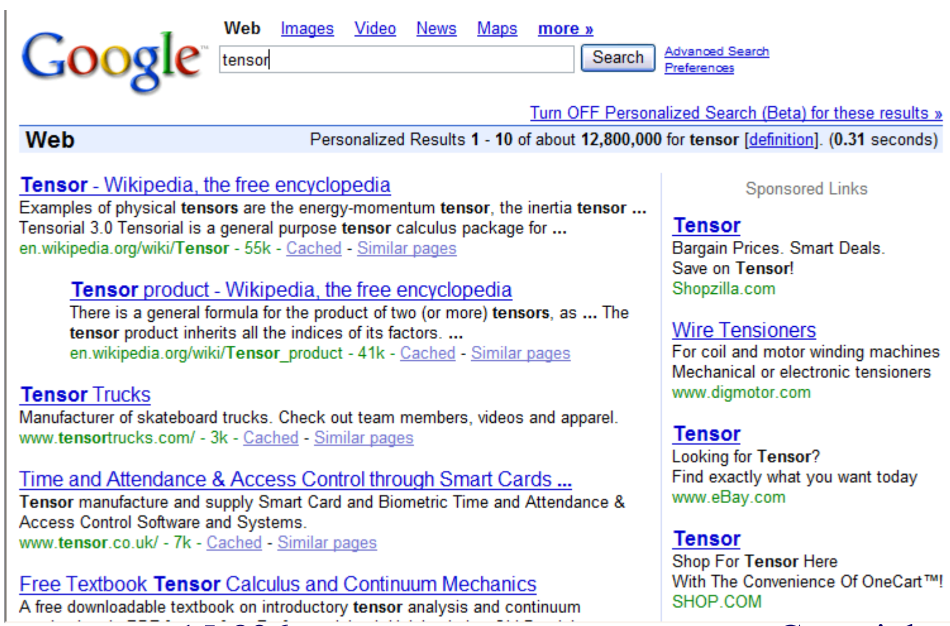
http://www.tensortoolbox.org/

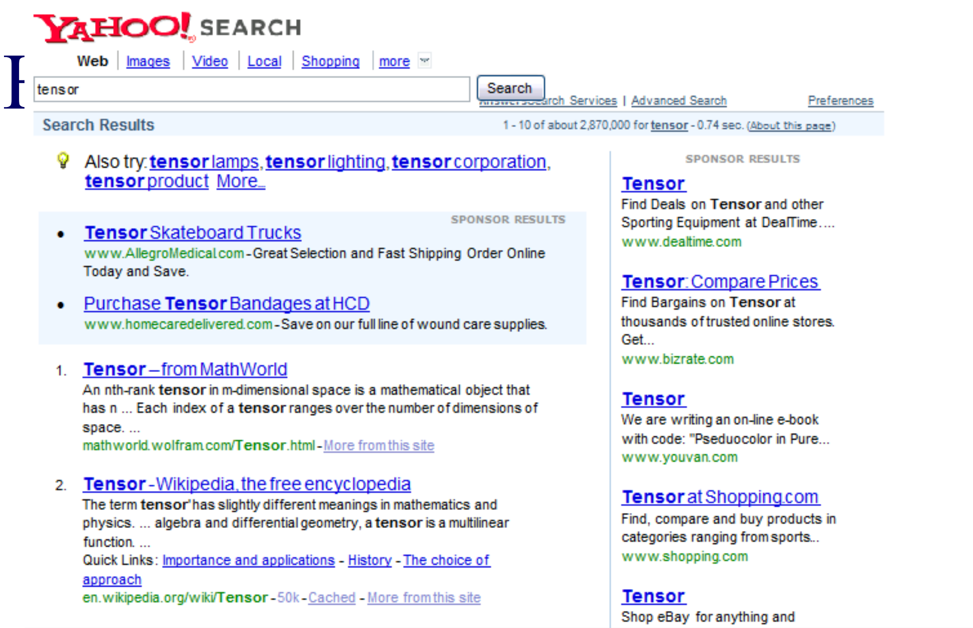
Outline

- Motivation - Definitions
- Tensor tools
- Case studies
  - P1: web graph mining (‘TOPHITS’)
  - P2: phone-call patterns
  - P3: N.E.L.L. (never ending language learner)
  - P4: network traffic
  - P5: FaceBook activity

P1: Web graph mining

- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank

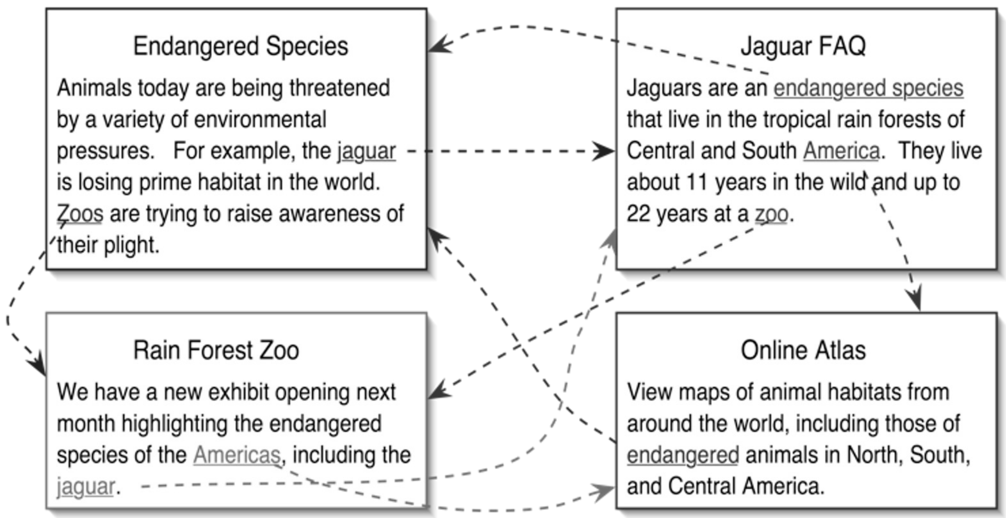




P1: Web graph mining

- T. G. Kolda, B. W. Bader and J. P. Kenny, ***Higher-Order Web Link Analysis Using Multilinear Algebra***, ICDM 2005: ICDM, pp. 242-249, November 2005, doi:10.1109/ICDM.2005.77. [PDF]

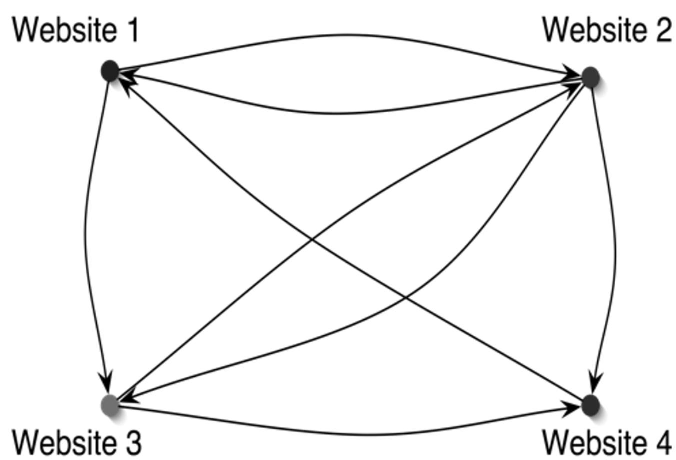
Kleinberg' s Hubs and Authorities (the HITS method)

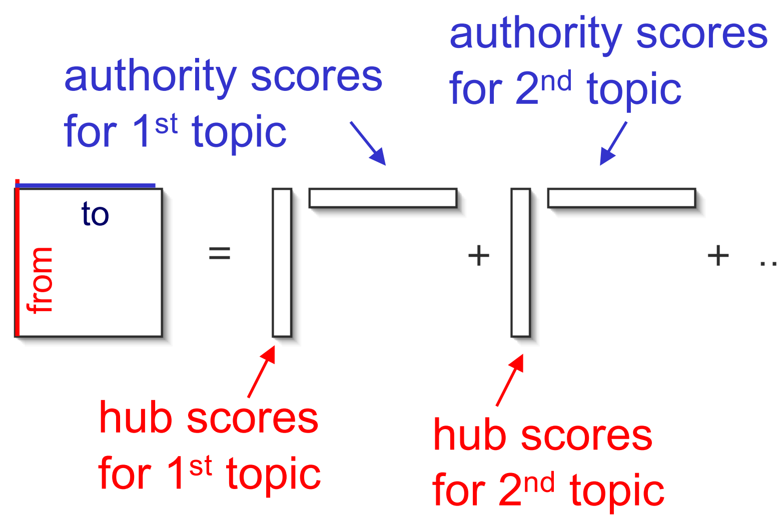


Sparse adjacency matrix and its SVD:

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

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





Kleinberg, JACM, 1999

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.co

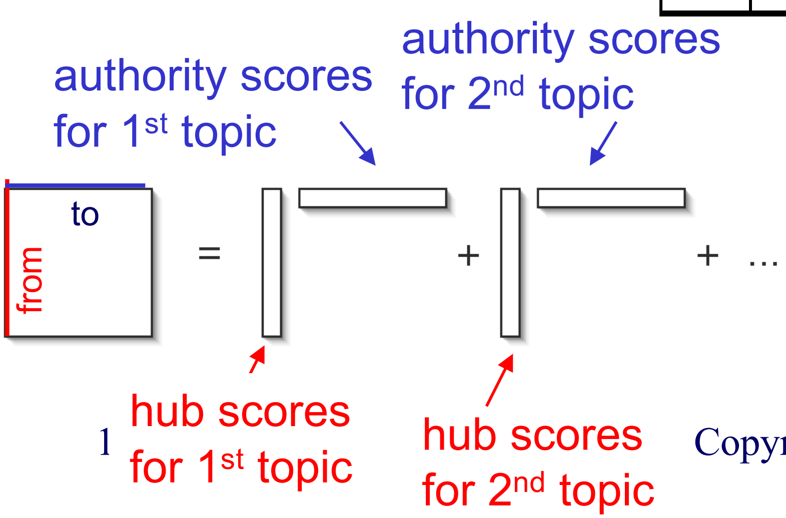
2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh
.06	www.lehigha
.06	www.lehighs
.02	www.bethleh
.02	www.adobe.c
.02	lewisweb.cc.
.02	www.leo.lehi
.02	www.distanc
.02	fp1.cc.lehigh

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-catalo
.08	news.com.com

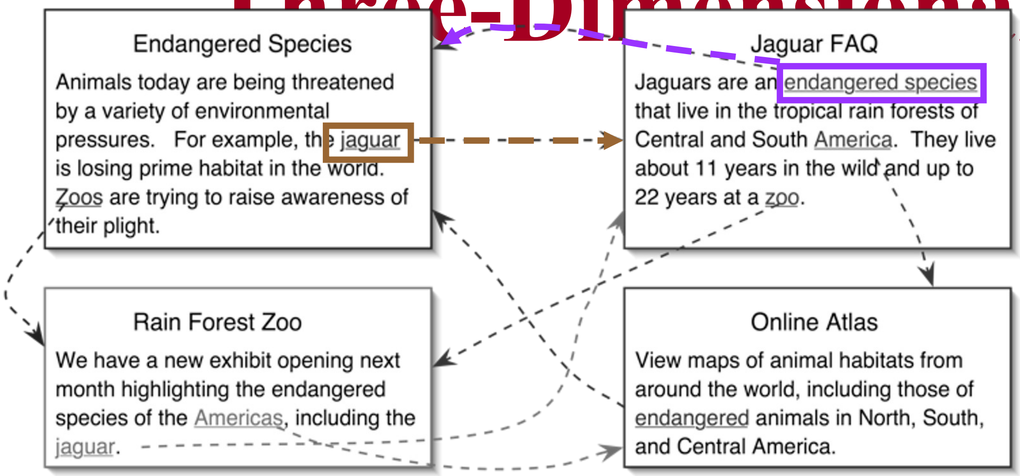
4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state
.22	www.gsa.g
.20	www.ssa.g
.16	www.censu
.14	www.govbe
.13	www.kids.g
.13	www.usdoj

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

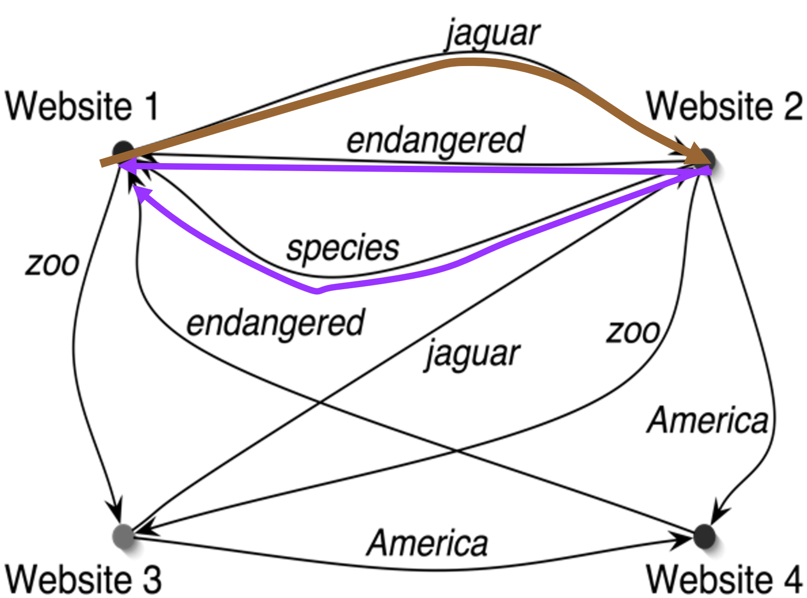
We started our crawl from http://www-neos.mcs.anl.gov/neos, and crawled 4700 pages, resulting in 560 cross-linked hosts.



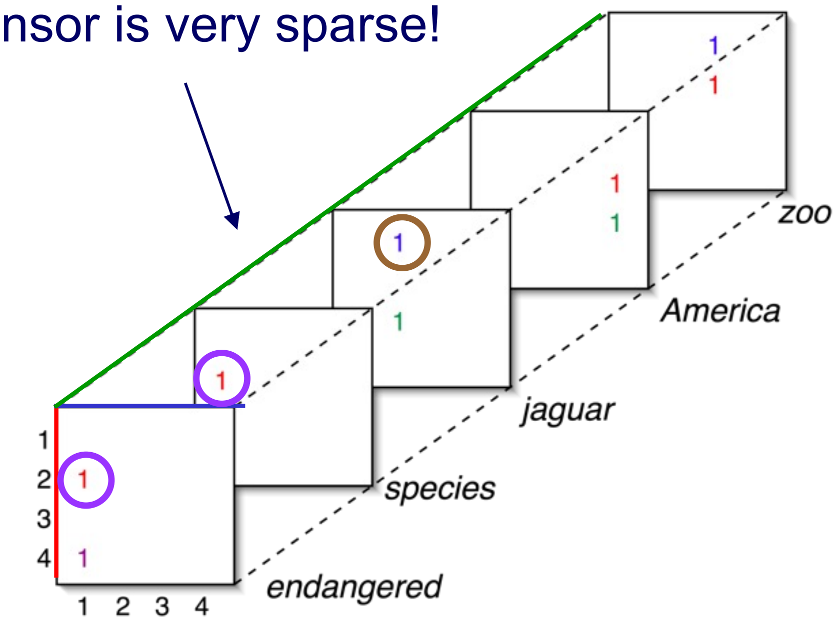
Three-Dimensional View of the Web



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



Observe that this tensor is very sparse!

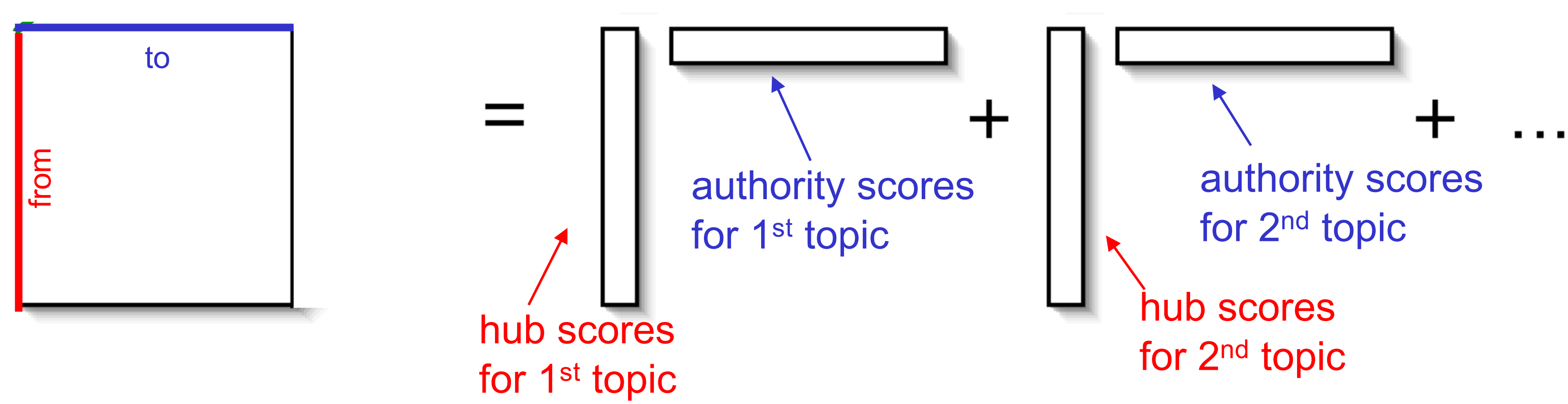


Kolda, Bader, Kenny, ICDM05

Topical HITS (TOPHITS)

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

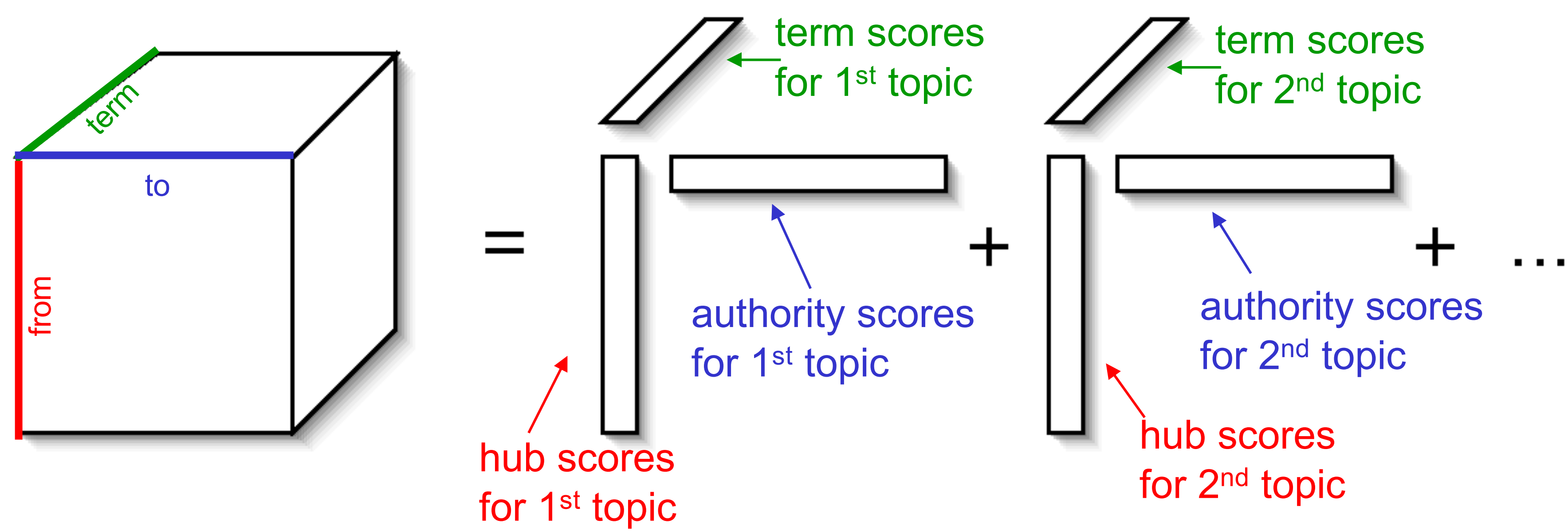
$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



Topical HITS (TOPHITS)

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

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



TOPHITS Terms & Authorities on Sample Data

1st Principal Factor

.23	JAVA	.86	java.sun.com
.18	SUN		
.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	LIBRA		
.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	CONS		
.13	FREE		

6th Principal Factor

.26	PRESIDENT	.87	www.whitehouse.gov
.25	NO-READABLE-TEXT	.18	www.irs.gov
.25	BUSH		
.25	WELC		
.17	WHIT		
.16	U.S		
.15	HOUS		
.13	BUDG		
.13	PRES		
.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	CONT		
.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

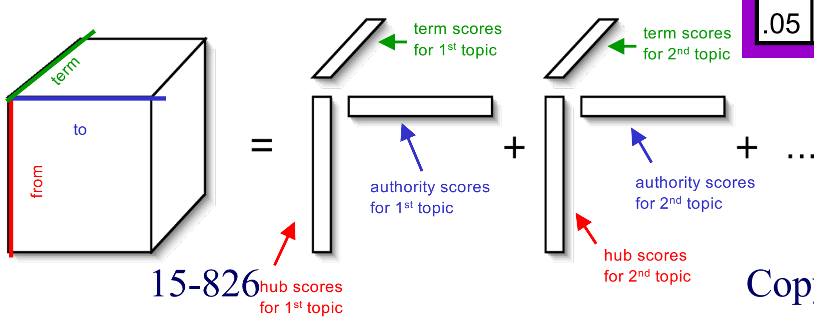
.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

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

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

w_k = # unique links using term k

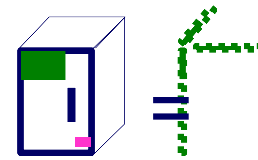
Tensor PARAFAC



Outline

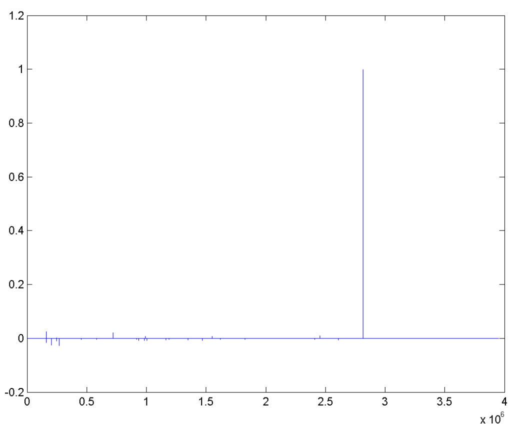
- Motivation - Definitions
- Tensor tools
- Case studies
  - P1: web graph mining ('TOPHITS')
  - P2: phone-call patterns
  - P3: N.E.L.L. (never ending language learner)
  - P4: network traffic
  - P5: FaceBook activity

P2: Anomaly detection in time-evolving graphs

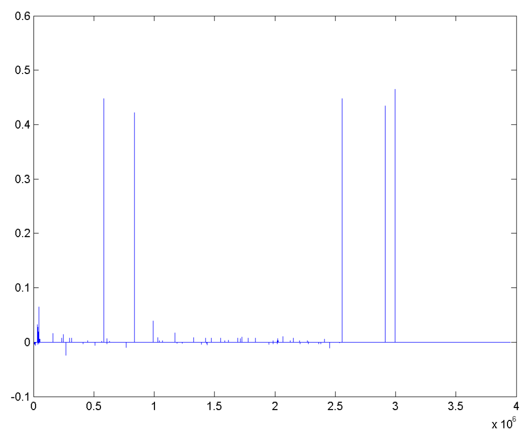


- Anomalous communities in phone call data:
  – European country, 4M clients, data over 2 weeks

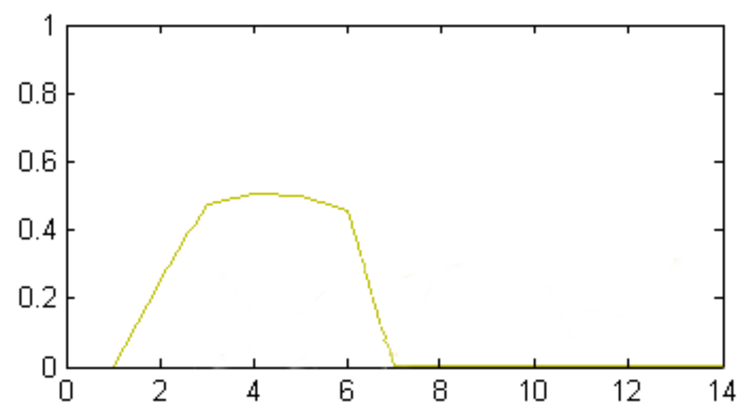
1 caller



5 receivers

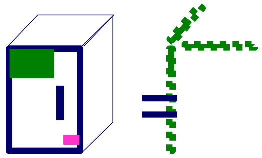


4 days of activity



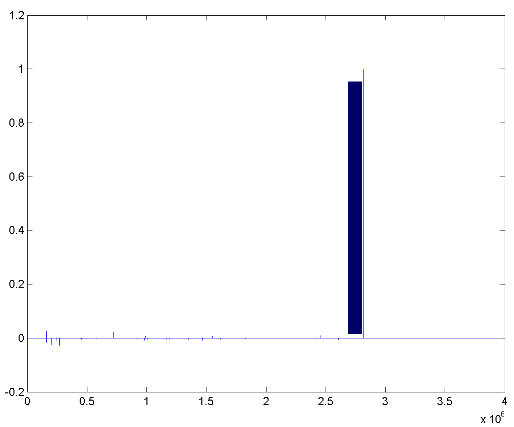
~200 calls to EACH receiver on EACH day!

P2: Anomaly detection in time-evolving graphs

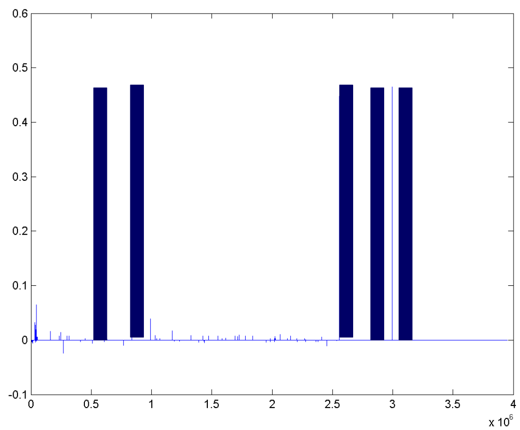


- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks

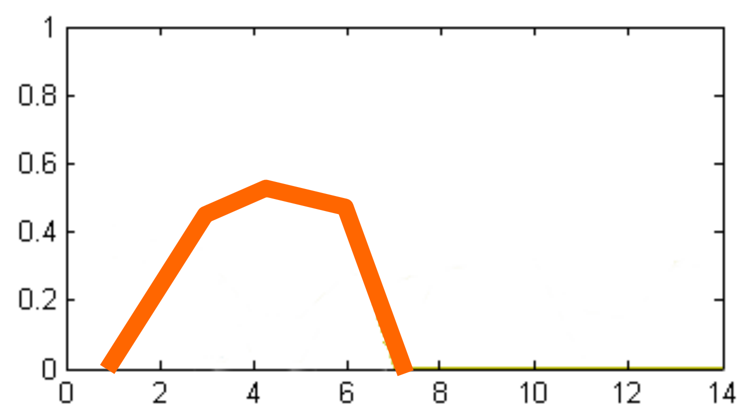
1 caller



5 receivers

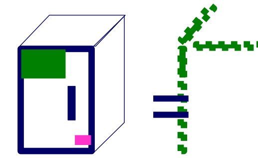


4 days of activity



~200 calls to EACH receiver on EACH day!

P2: Anomaly detection in time-evolving graphs

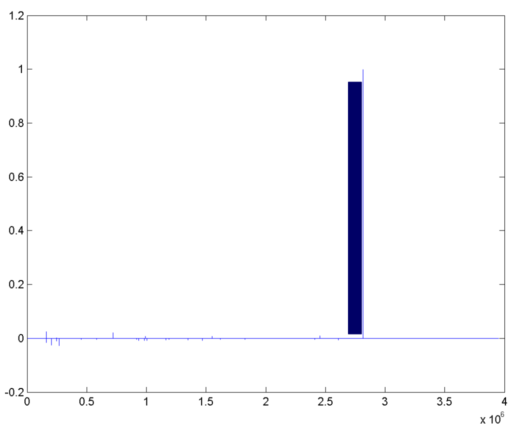


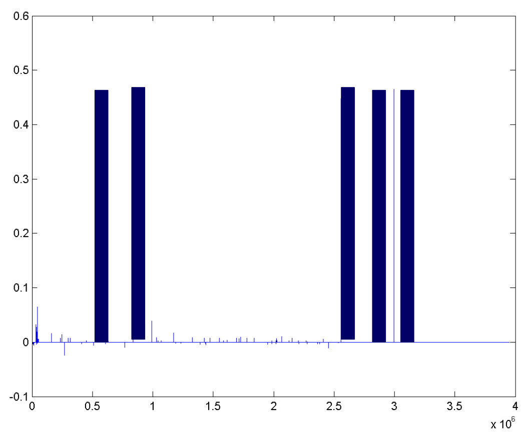
- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks

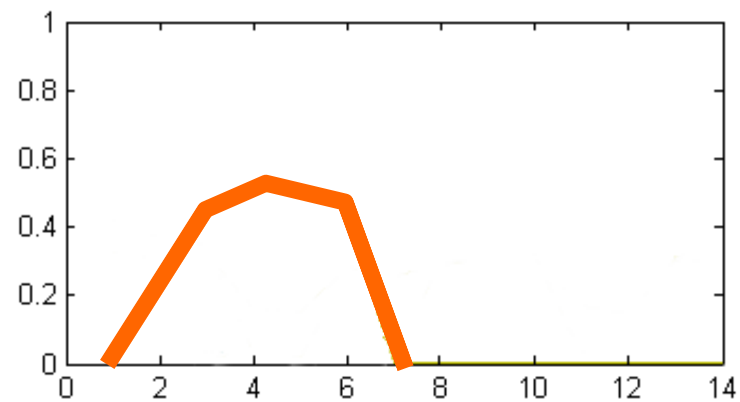
1 caller

5 receivers

4 days of activity

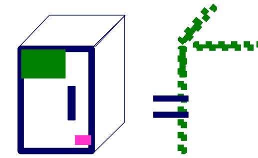






~200 calls to EACH receiver on EACH day!

P2: Anomaly detection in time-evolving graphs



- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks

Miguel Araujo, Spiros Papadimitriou, Stephan Günnemann, Christos Faloutsos, Prithwish Basu, Ananthram Swami, Evangelos Papalexakis, Danai Koutra. *Com2: Fast Automatic Discovery of Temporal (Comet) Communities*. PAKDD 2014, Tainan, Taiwan.

Carnegie Mellon

Outline

- Motivation - Definitions
- Tensor tools
- Case studies
  - P1: web graph mining (‘TOPHITS’)
  - P2: phone-call patterns
  - P3: N.E.L.L. (never ending language learner)
  - P4: network traffic
  - P5: FaceBook activity

GigaTensor: Scaling Tensor Analysis Up By 100 Times – Algorithms and Discoveries

U Kang | Evangelos Papalexakis | Abhay Harpale | Christos Faloutsos

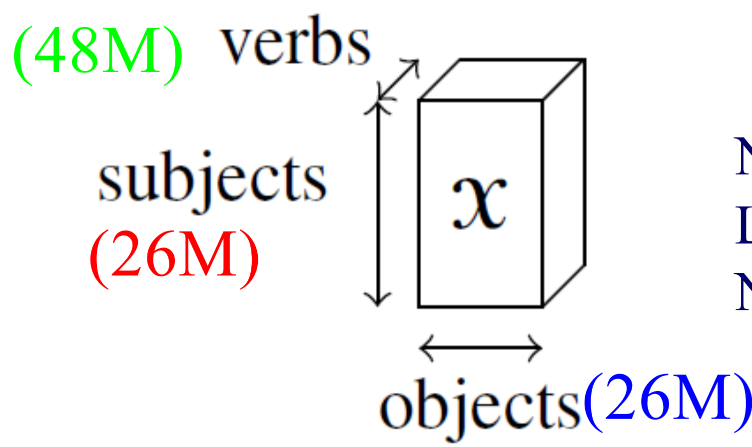
KDD 2012

P3: 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"

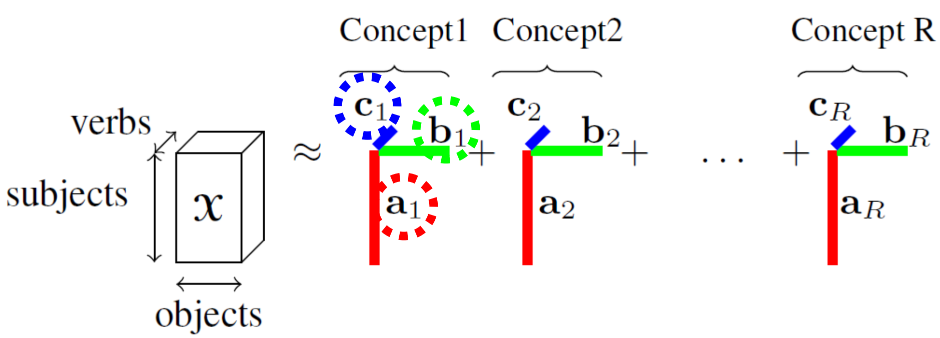
"Barrack Obama is the president of U.S."



NELL (Never Ending Language Learner)
Nonzeros =144M

A1: Concept Discovery

- Concept Discovery in Knowledge Base



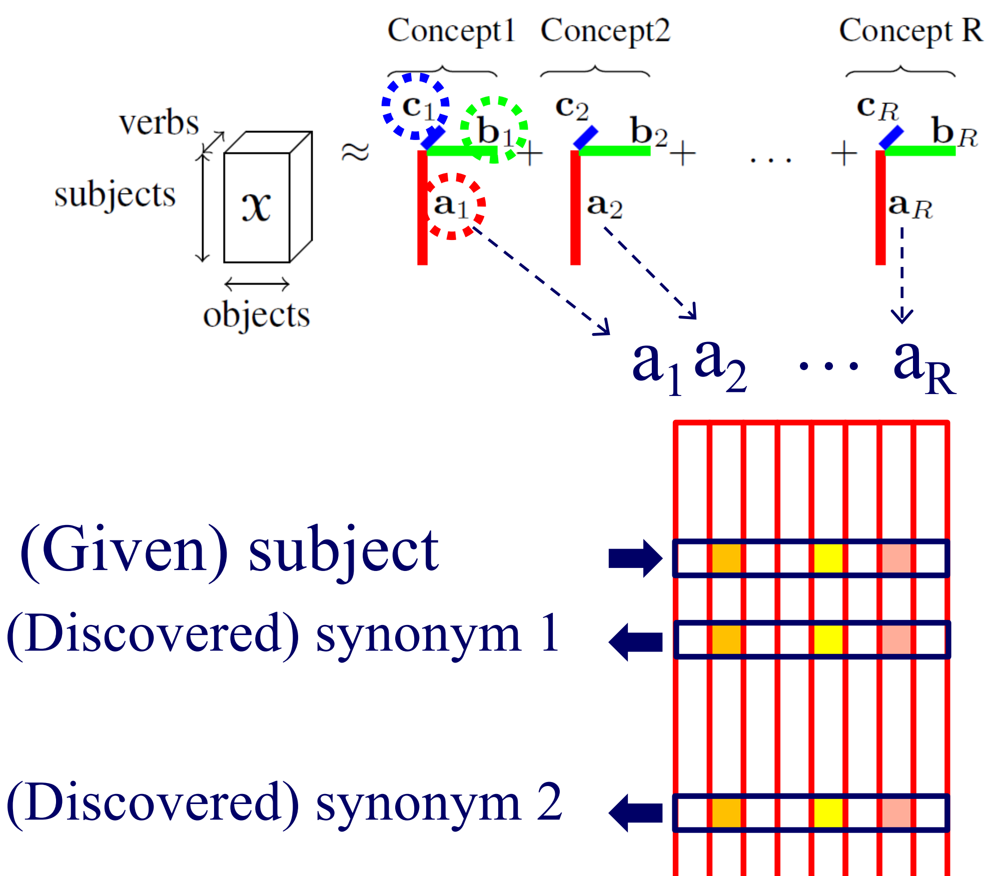
Noun Phrase 1	Noun Phrase 2	Context
Concept 1: "Web Protocol"		
internet	protocol	‘np1’ ‘stream’ ‘np2’
file	software	‘np1’ ‘marketing’ ‘np2’
data	suite	‘np1’ ‘dating’ ‘np2’
Concept 2: "Credit Cards"		
credit	information	‘np1’ ‘card’ ‘np2’
Credit	debt	‘np1’ ‘report’ ‘np2’
library	number	‘np1’ ‘cards’ ‘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
  - P1: web graph mining (‘TOPHITS’)
  - P2: phone-call patterns
  - P3: N.E.L.L. (never ending language learner)
  - P4: network traffic
  - P5: FaceBook activity

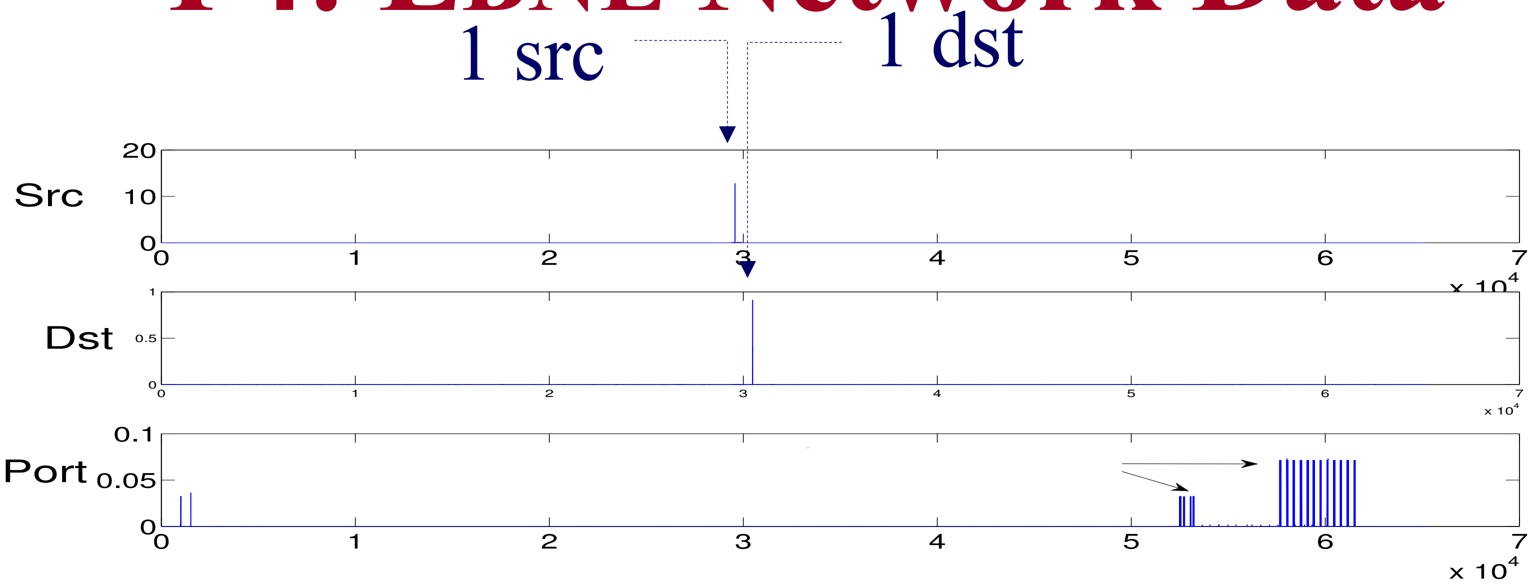
Carnegie Mellon

ParCube: Sparse Parallelizable Tensor Decompositions

Evangelos E. Papalexakis, Christos Faloutsos, Nikos Sidiropoulos,
ECML/PKDD 2012

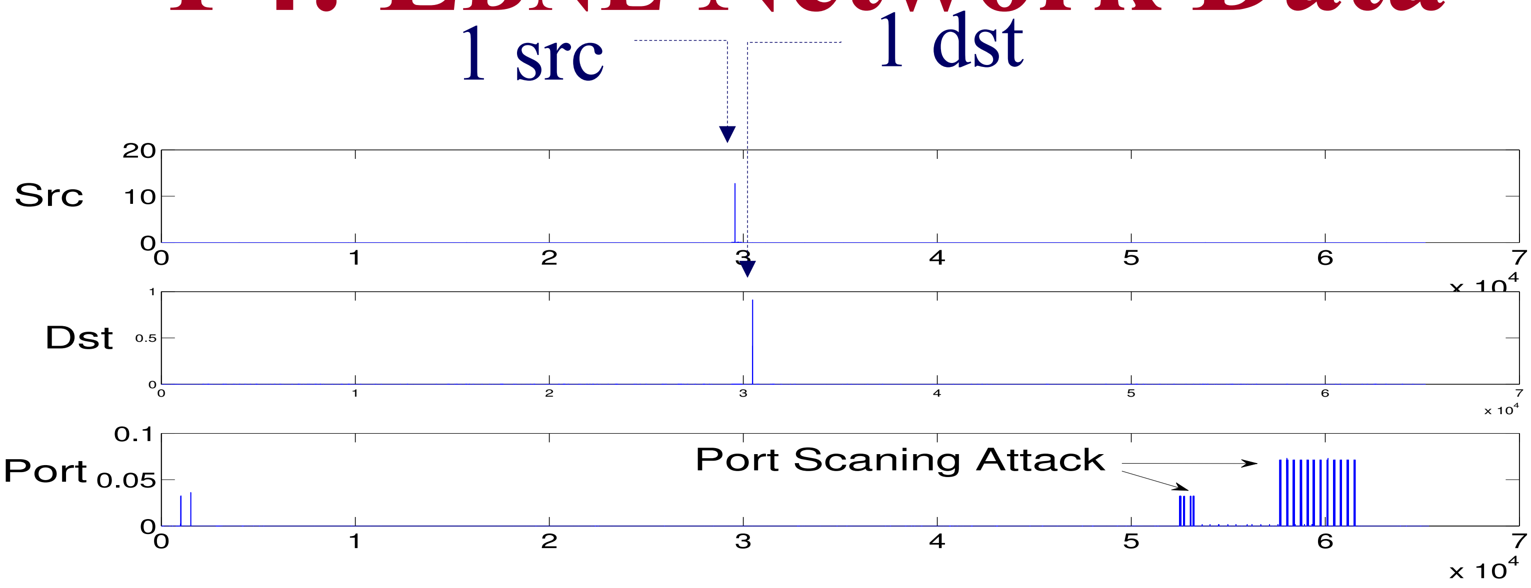
Evangelos E. Papalexakis
Email: epapalex@cs.ucr.edu
Web: http://www.cs.ucr.edu/~epapalex

P4: LBNL Network Data



- Modes: src IP, dst IP, port #
- Q: what could it be?
- **A:**

P4: LBNL Network Data

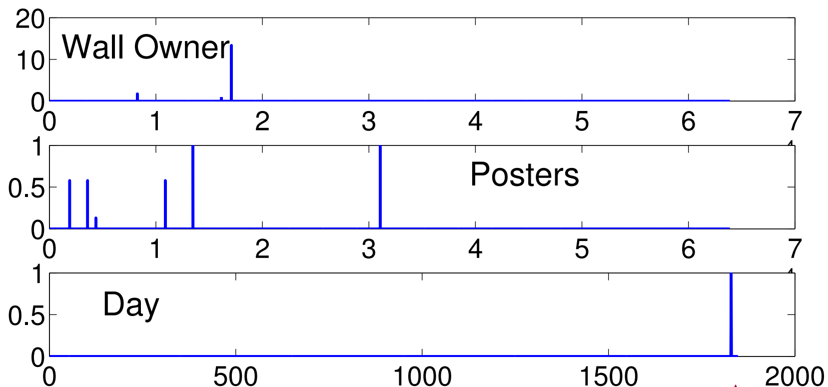


- Modes: src IP, dst IP, port #
- Q: what could it be?
- **A: ~ Port Scanning Attack**

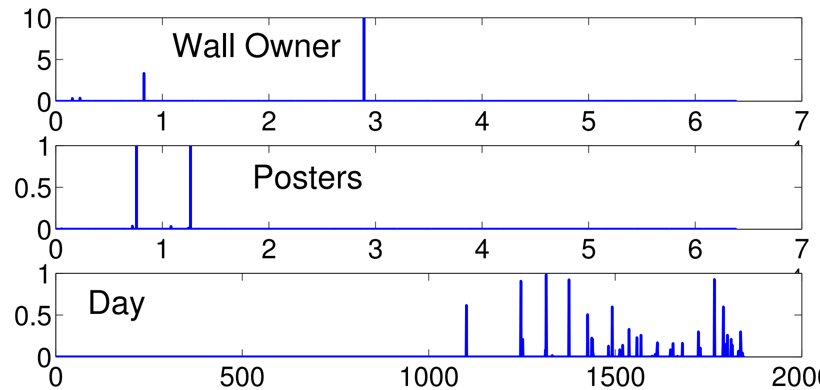
Carnegie Mellon

P5: FACEBOOK Wall posts

1 Wall



(a) FACEBOOK anomaly



(b) FACEBOOK normal activity

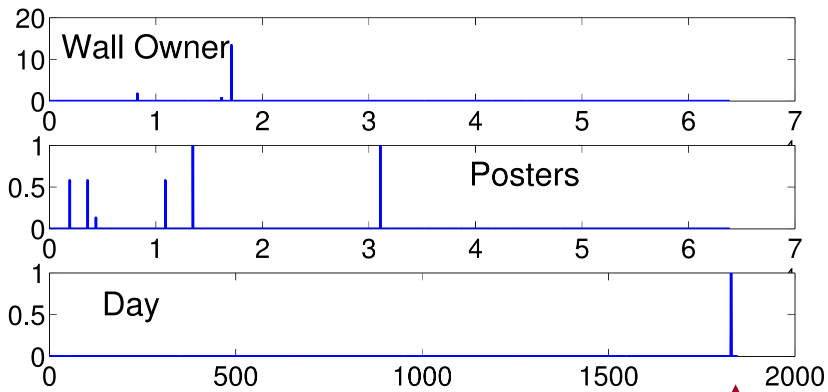
1 day

- Modes: wall-owner, poster, timestamp
- Discovery: What could it be?

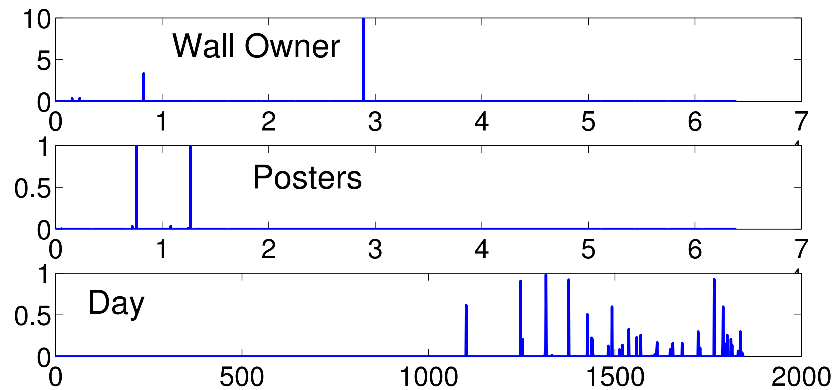
Carnegie Mellon

P5: FACEBOOK Wall posts

1 Wall



(a) FACEBOOK anomaly (Wall owner's birthday)

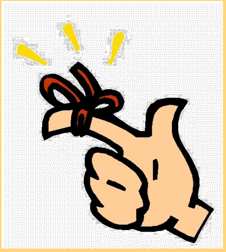


(b) FACEBOOK normal activity

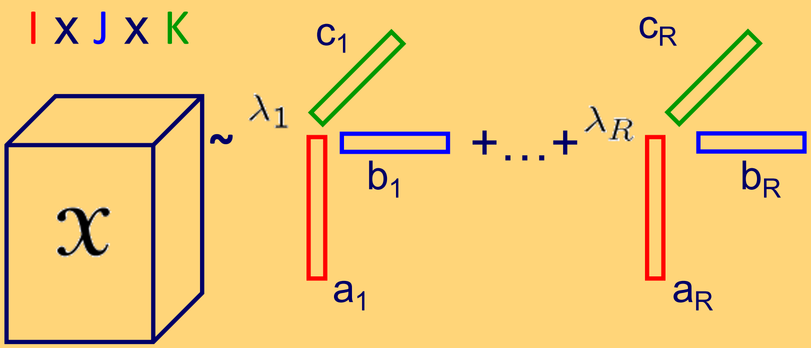
1 day

- Modes: wall-owner, poster, timestamp
- Discovery: What could it be? A: Birthday

Conclusions



- Q: who-calls-whom-when – patterns?
  - Triplets (source-ip, dest-ip, port#)
  - KB (subject, verb, object)
- A: Tensor analysis (PARAFAC)
  - http://www.tensortoolbox.org/



References

- Inderjit S. Dhillon, Subramanyam Mallela, Dharmendra S. Modha: Information-theoretic co-clustering. KDD 2003: 89-98
- 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

Carnegie Mellon

References

- Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, Christos Faloutsos: *Tensor Decomposition for Signal Processing and Machine Learning*. IEEE Trans. Signal Process. 65(13): 3551-3582 (2017)

Software

- Tensorly: http://tensorly.org/stable/index.html
- Tensor toolbox (in matlab) https://www.tensortoolbox.org/

