Mining Large Time-evolving Data Using Matrix and Tensor Tools

Christos Faloutsos Carnegie Mellon Univ.
Tamara G. Kolda Sandia National Labs
Jimeng Sun Carnegie Mellon Univ.

About the tutorial

- Introduce **matrix and tensor tools** through **real mining applications**

- **Goal:** find **patterns, rules, clusters, outliers, …**
  - in matrices and
  - in tensors
- www.cs.cmu.edu/~christos/TALKS/SIGMOD-07-tutorial/

What is this tutorial about?

- Matrix tools
  - Singular Value Decomposition (SVD)
  - Principal Component Analysis (PCA)
  - Webpage ranking algorithms: HITS, PageRank
  - CUR decomposition
  - Co-clustering
- Tensor tools
  - Tucker decomposition
  - Parallel factor analysis (PARAFAC)
  - Incrementalization
- Applications

What is this tutorial NOT about?

- Classification methods
- Kernel methods
- Discriminative models
  - Linear Discriminant Analysis (LDA)
  - Canonical Correlation Analysis (CCA)
- Probabilistic latent variable models
  - Probabilistic PCA
  - Probabilistic latent semantic indexing
  - Latent Dirichlet allocation

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 2: Why tensors?

- Q: what is a tensor?

Motivation 2: Why tensor?

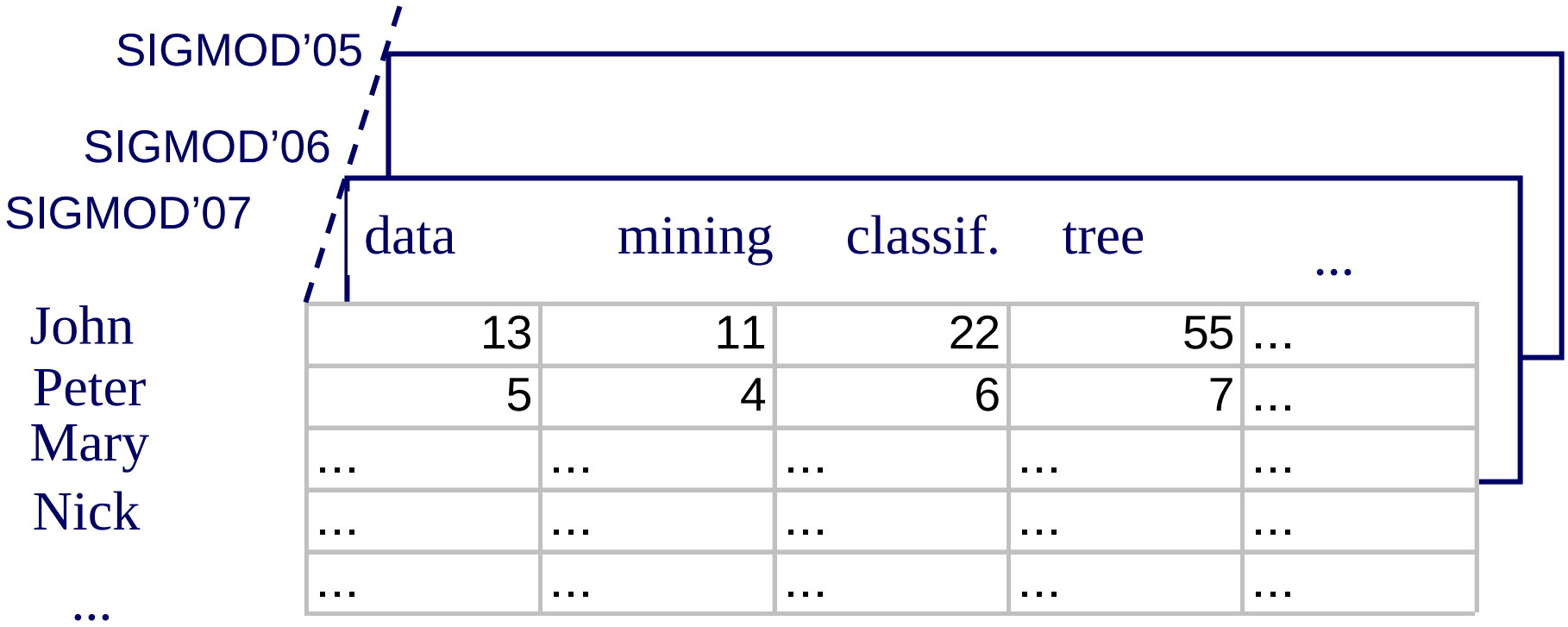
- A: N-D generalization of matrix:

SIGMOD'07

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

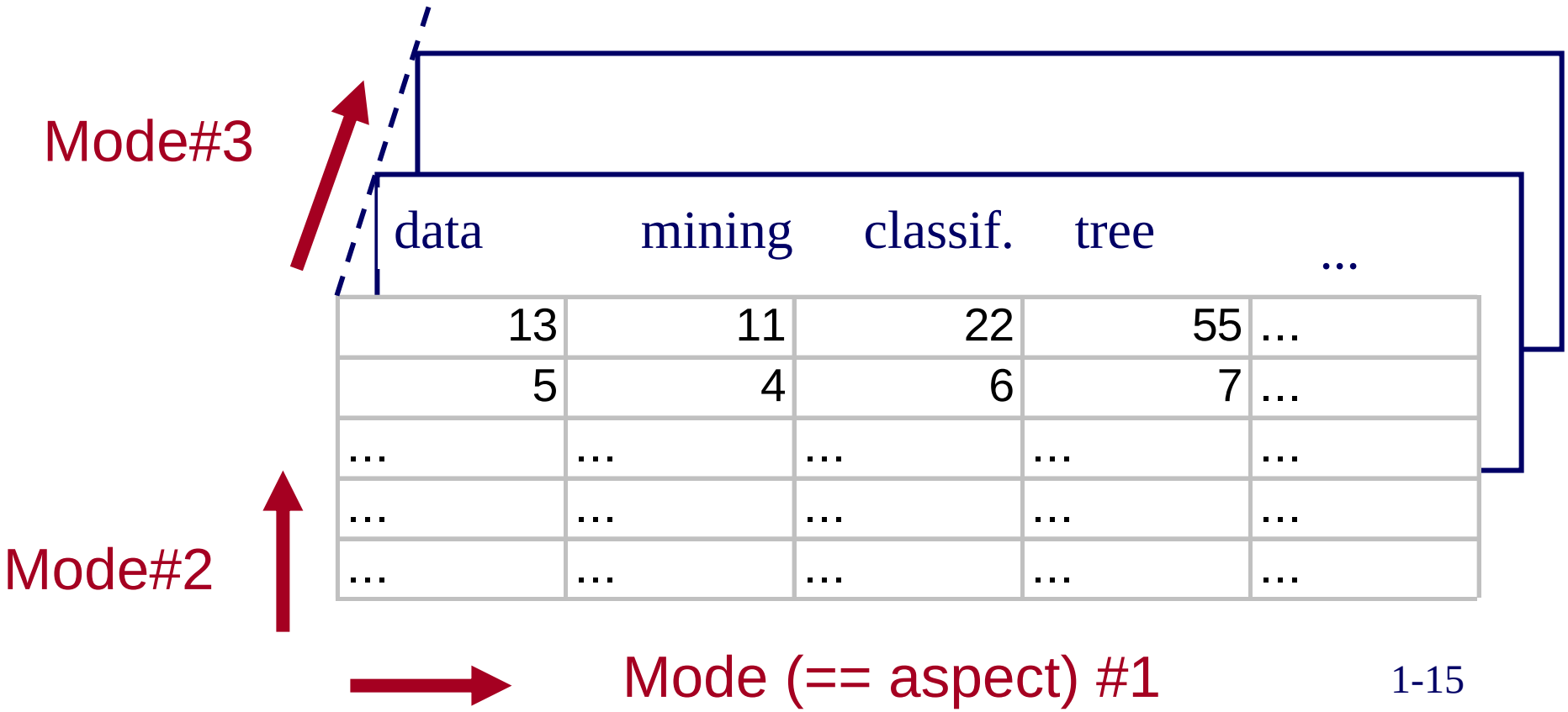
Motivation 2: Why tensor?

- A: N-D generalization of matrix:



Tensors are useful for 3 or more modes

Terminology: ‘mode’ (or ‘aspect’):



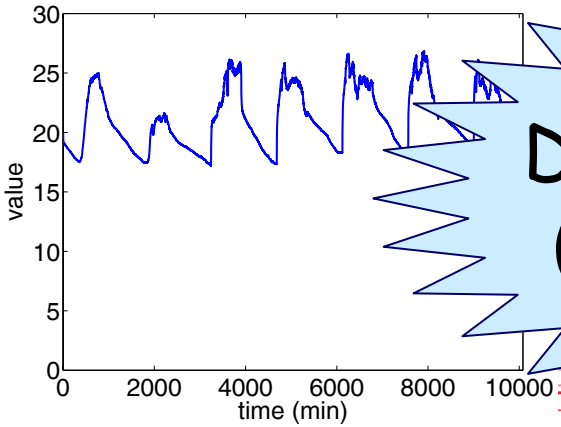
Motivating Applications

- Why matrices are important?
- Why tensors are useful?
  - P1: environmental sensors
  - P2: data center monitoring (‘autonomic’)
  - P3: social networks
  - P4: network forensics
  - P5: web mining
  - P6: face recognition

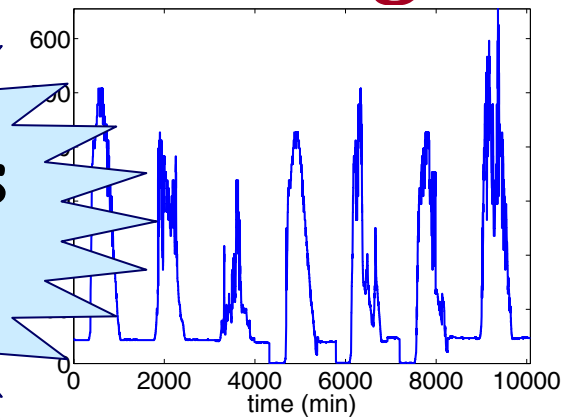
Sandia National Laboratories

P1: Environmental sensor monitoring

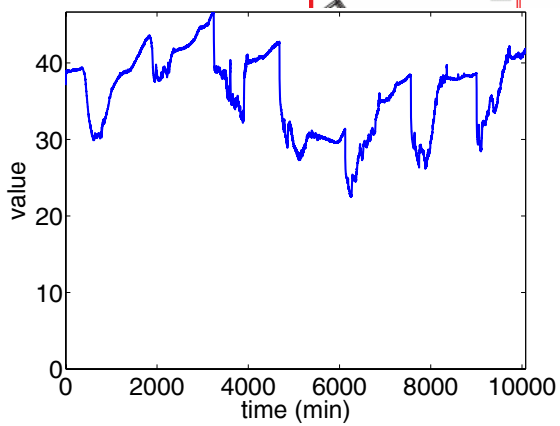
Data in three **aspects** (time, location, type)



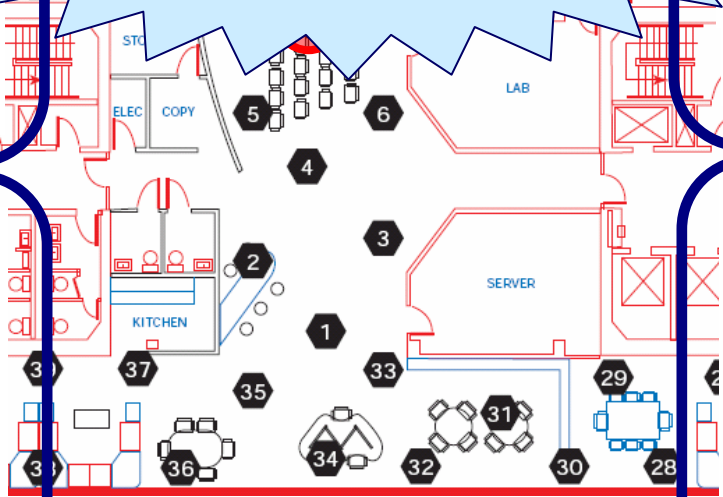
Temperature

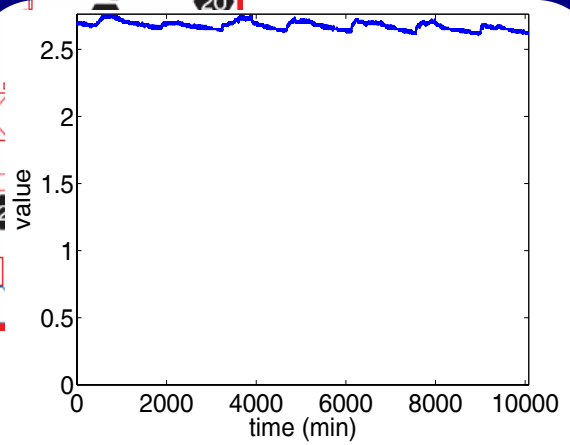


Light



Humidity

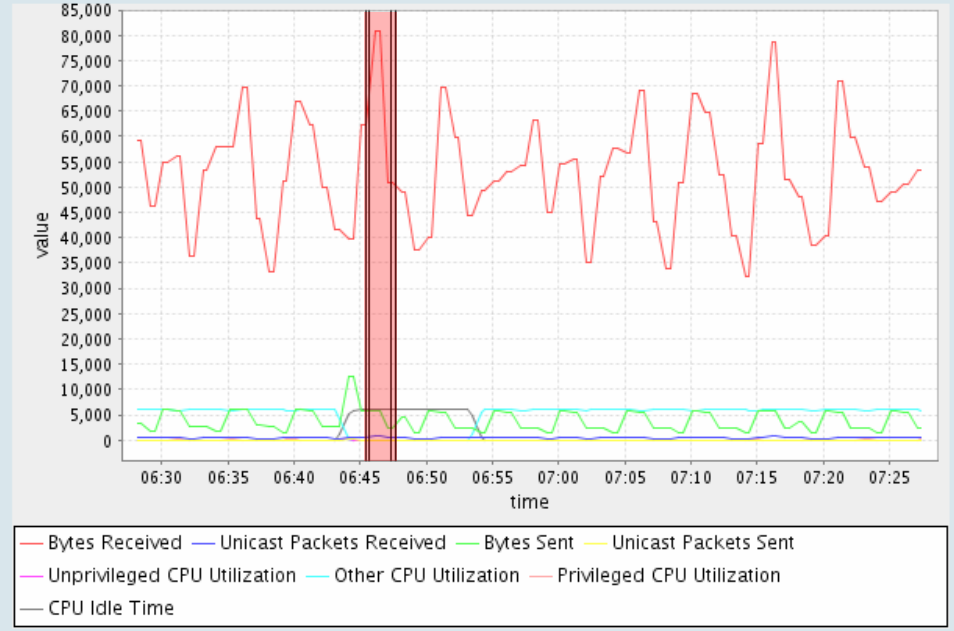


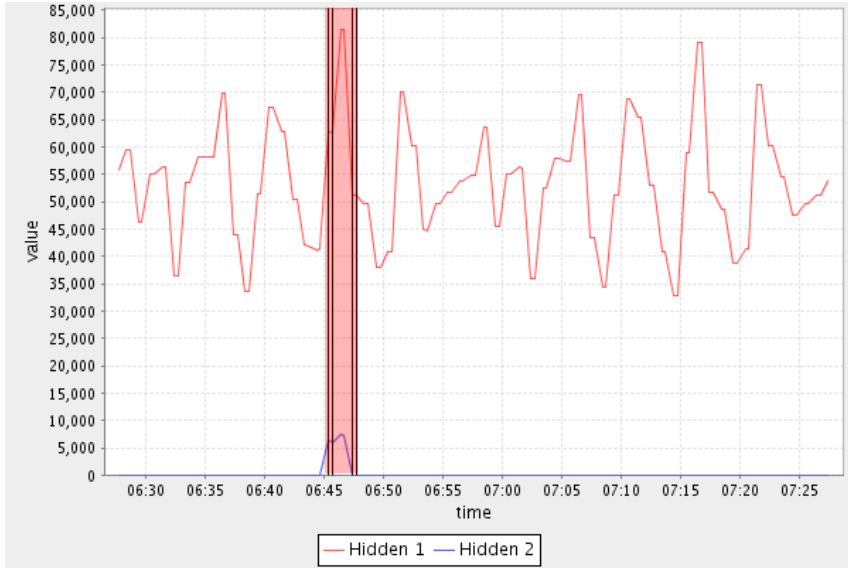


Voltage

Sandia National Laboratories

P2: Clusters/data center monitoring

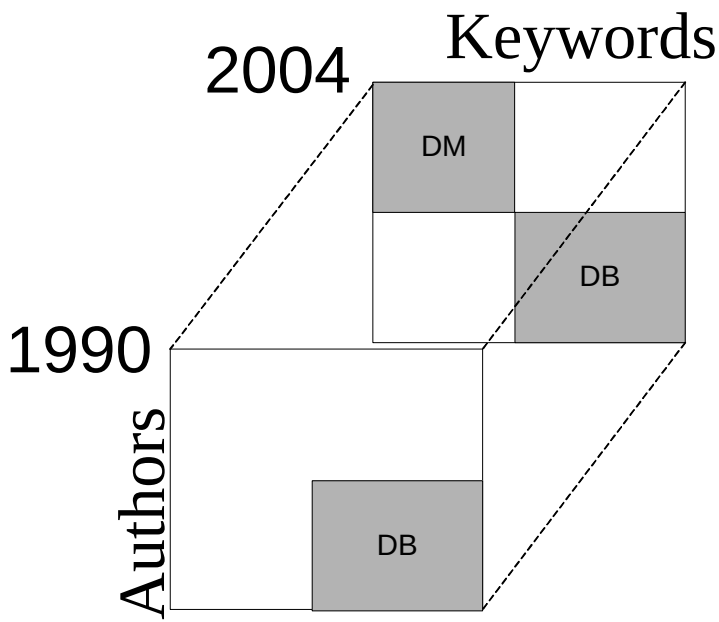




- Monitor correlations of multiple measurements
- Automatically flag anomalous behavior
- Intemon: intelligent monitoring system
  - Prof. Greg Ganger and PDL
  - >100 machines in a data center
  - warsteiner.db.cs.cmu.edu/demo/intemon.jsp

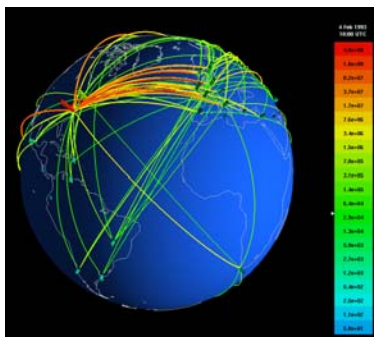
P3: Social network analysis

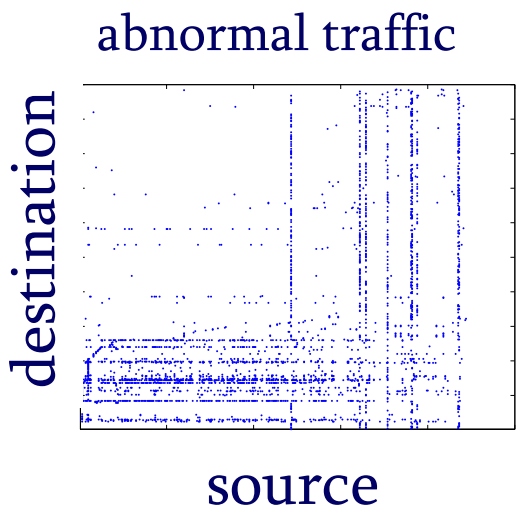
- Traditionally, people focus on static networks and find community structures
- We plan to monitor the change of the community structure over time

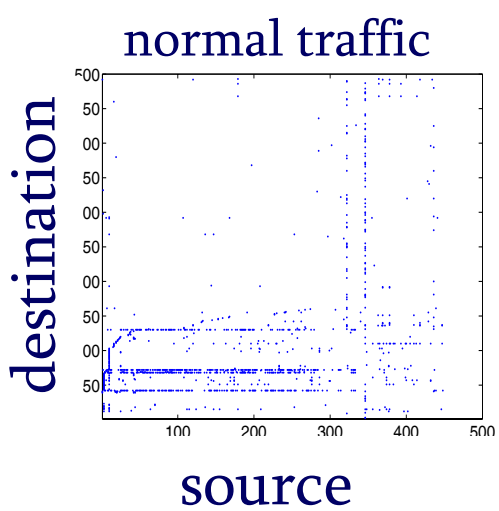


P4: Network forensics

- Directional network flows
- A large ISP with 100 POPs, each POP 10Gbps link capacity [Hotnets2004]
  - 450 GB/hour with compression
- Task: Identify abnormal traffic pattern and find out the cause







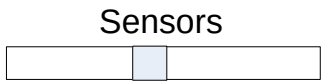
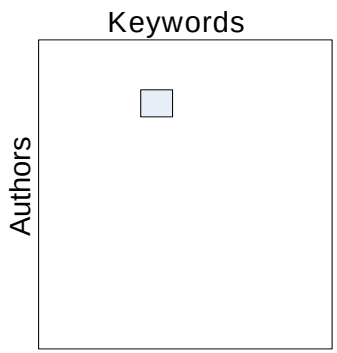
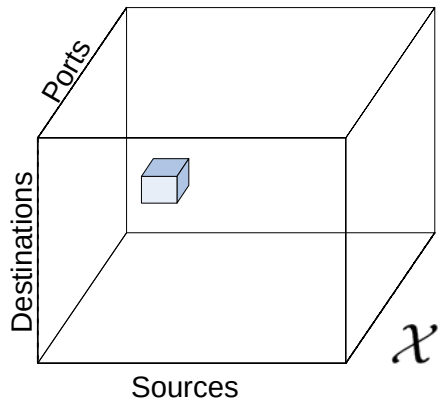
Collaboration with Prof. Hui Zhang and Dr. Yinglian Xie

P5: 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

Static Data model

- Tensor
  - Formally, $\mathcal{X} \in \mathbf{R}^{N_1 \times \ldots \times N_M}$
  - Generalization of matrices
  - Represented as multi-array, (~ data cube).

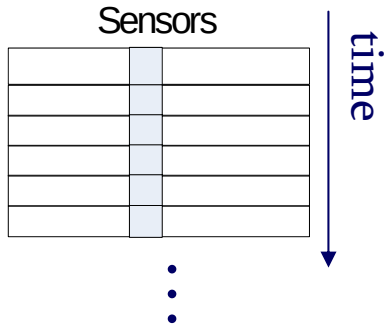
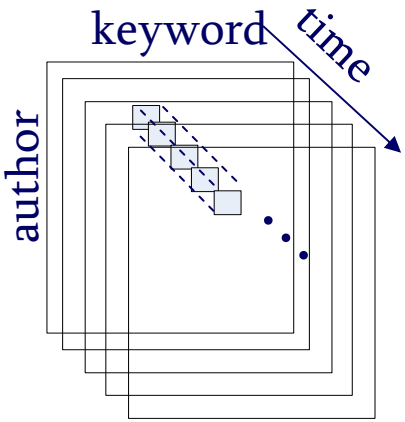
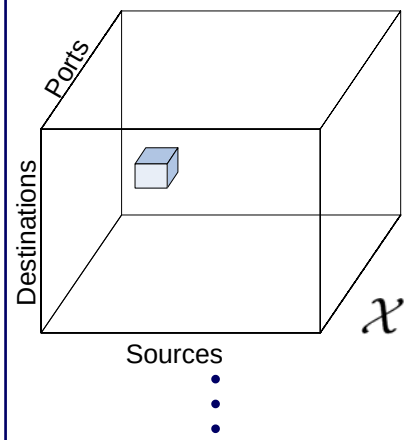
Order	1st	2nd	3rd
Correspondence	Vector	Matrix	3D array
Example			

Dynamic Data model

- Tensor Streams
  - A sequence of Mth order tensors

$$\mathcal{X}_1 \ldots \mathcal{X}_t \text{ where } \mathcal{X}_i \in \mathbf{R}^{N_1 \times \ldots \times N_M}$$

t is increasing over time

Order	1st	2nd	3rd
Correspondence	Multiple streams	Time evolving graphs	3D arrays
Example			

Roadmap

- Motivation
- Matrix tools
- Tensor tools
- Case studies