School of Computer Science
Carnegie Mellon University

TENSORSPLAT: Spotting Latent Anomalies in Time

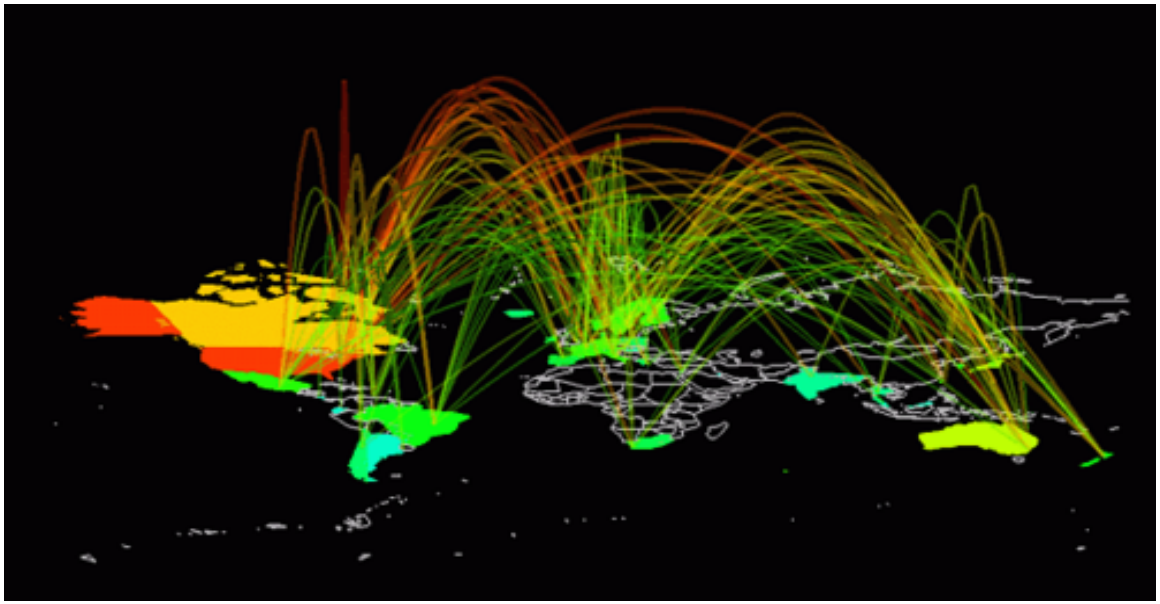
Danai Koutra, Evangelos E. Papalexakis, Christos Faloutsos
School of Computer Science, Carnegie Mellon University

*Kindly presented by **Dimitra Tzanetatou** and **Kostas Apostolou***

Panhellenic Conference on Informatics (PCI), October 5- 7, 2012 University of Piraeus, Greece

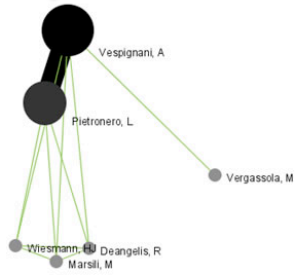
Motivation: Network Traffic

- Data:
  - which ***source IP*** contacted what ***destination IP***, on what ***Port #*** and ***when***
- How can we find possible network attacks on this, potentially large scale data?

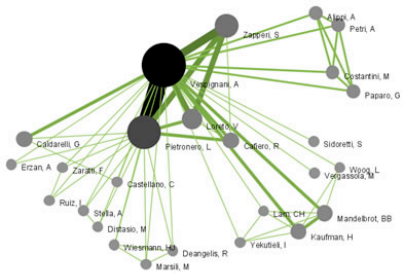


Motivation: Citation Network

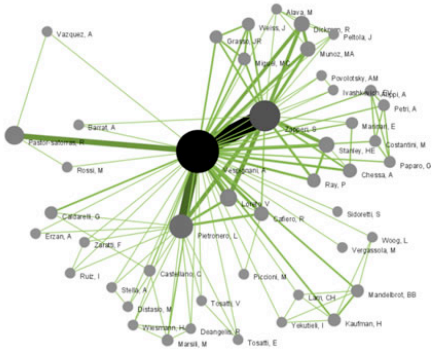
- DBLP data:
  - ***Who*** publishes ***what*** in ***which*** conference
  - In ***which*** conferences an ***author*** publishes **every** *year*
- How can we automatically cluster authors with similar research interests?
- How can we spot **"bridge authors"** who at some point change fields?



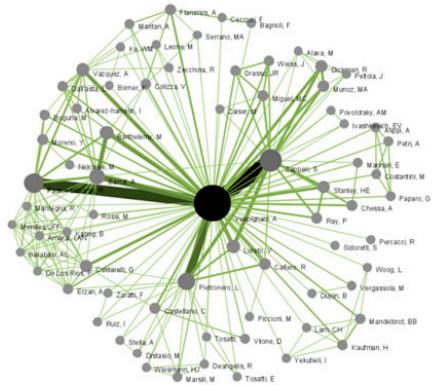
1990-1991



1990-1996



1990-2001

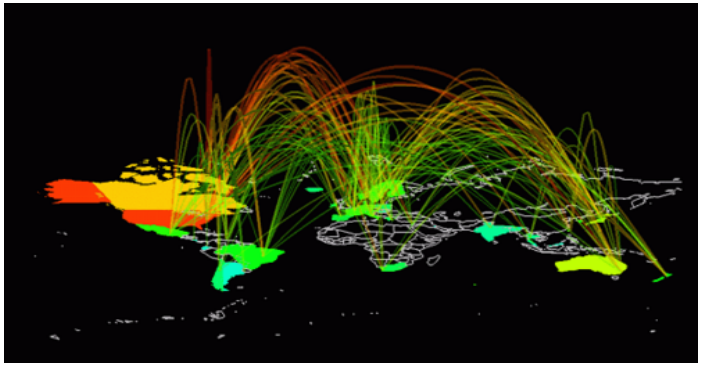


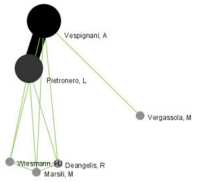
1990-2006

Motivation: Social Networks

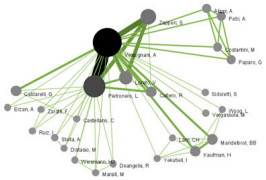
- Facebook Data (~800 Million users)
  - ***who*** posted on ***what*** *wall* and ***when***.
- How do we spot interesting patterns & anomalies in this very large network?

How to answer these questions?

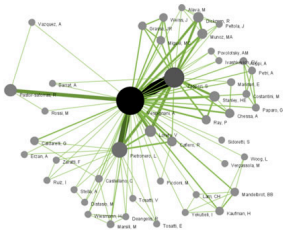




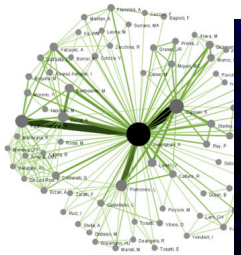
1990-1991



1990-1996



1990-2001



1990-2006

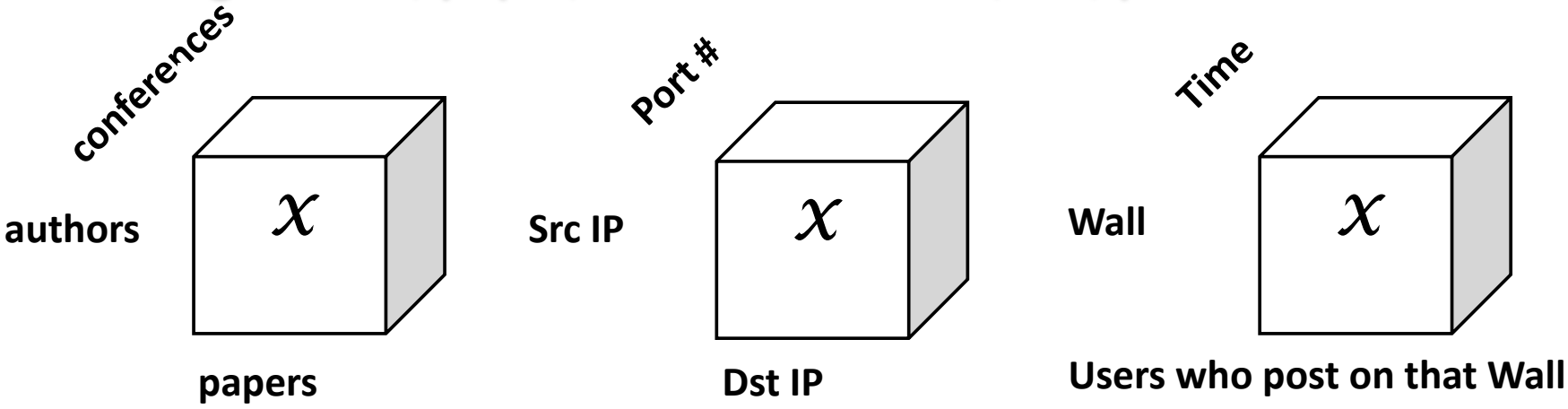
Our approach: a powerful tool called **TENSORS**

Outline

- **Introduction to Tensors**
- Applications
- Conclusions

Introduction to Tensors (1)

- One answer to the previous problems is Tensors!
- Tensors are multidimensional generalizations of matrices
  - A 3-way tensor is a 3-dimensional matrix or "cube"
- Lots of data can be modeled as a tensor:
  - Time-evolving graphs/social networks, Multi-aspect data e.g. author, paper, conference or src, dst, port#

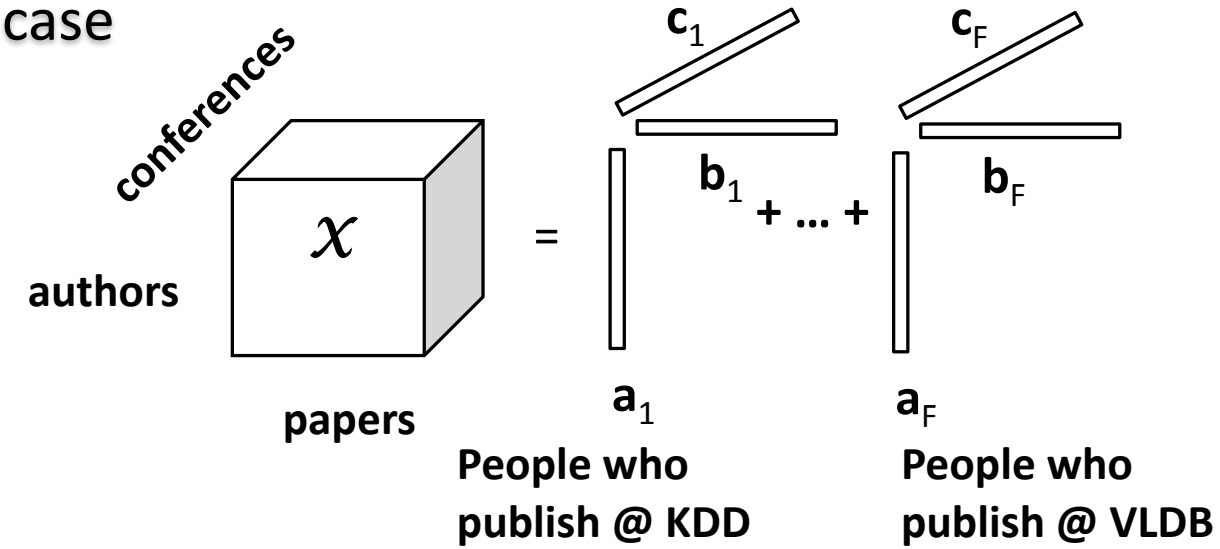


Introduction to Tensors (2)

- In the previous slide, we showed examples of 3-way tensors.
- Can have higher ways too!
  - E.g Network traffic data is in fact 4-way:
    - Src IP, Dst IP, Port # , Timestamp
- Harder to visualize on paper
  - But same principles apply
  - Same kind of analysis!

Introduction to Tensors (3)

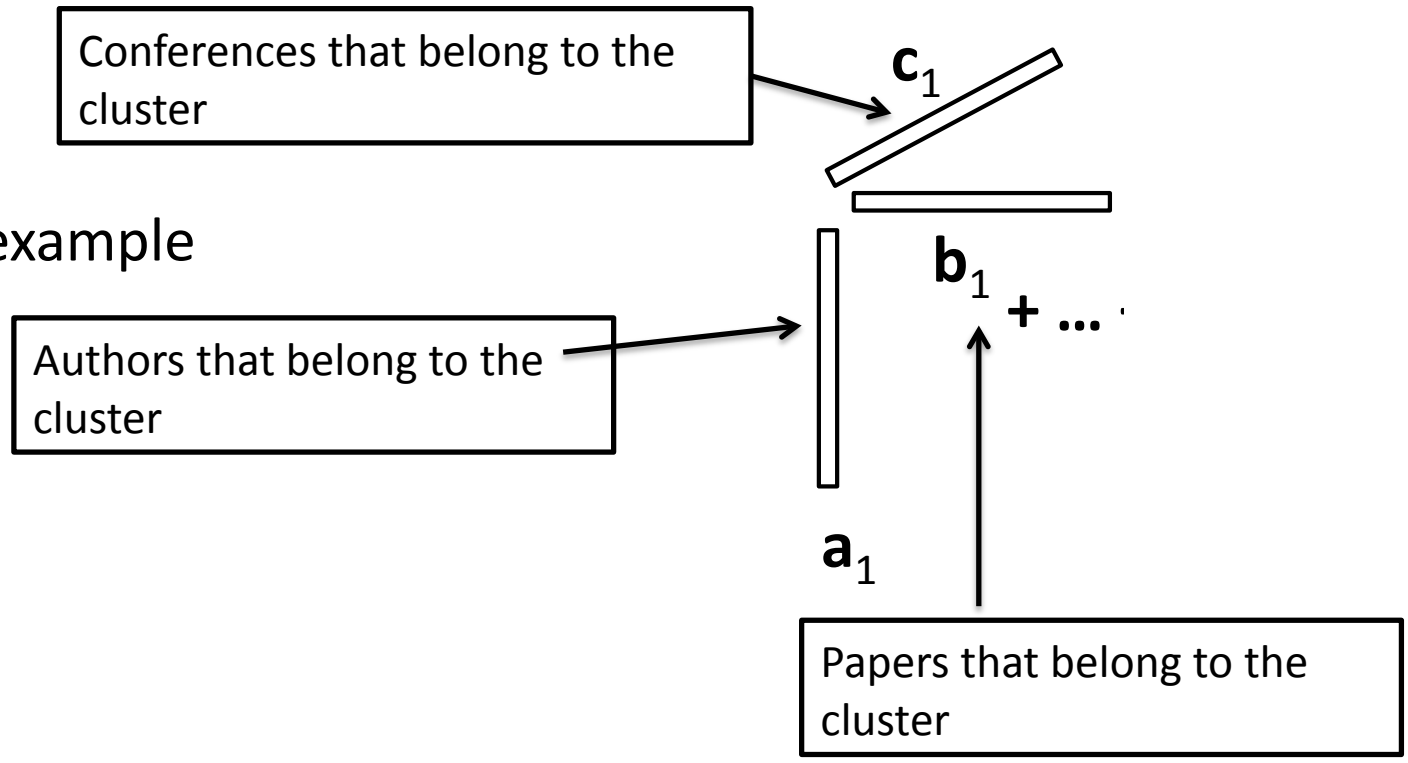
- PARAFAC decomposition
  - Decompose a tensor into sum of outer products/rank 1 tensors
  - Each rank 1 tensor is a different group/"concept"
    - This way, we can do *soft* clustering!
  - "Similar" to the Singular Value Decomposition in the matrix case



Introduction to Tensors (4)

- Each **a**, **b, c** triplet can be seen as "soft" membership to a cluster
- If we have 4-way tensor (e.g. Network Traffic), we have a fourth vector **d**

In the previous example (author, paper, conference):



Data Analysis

- We use PARAFAC with Non-negativity (NN) constraints
  - NN is important for interpretation (soft clustering membership can't be negative)
- We use the *Tensor Toolbox for Matlab* which is able to handle efficiently tensors with sparse representation
  - *Download at: http://www.sandia.gov/~tgkolda/TensorToolbox/index-2.5.html*

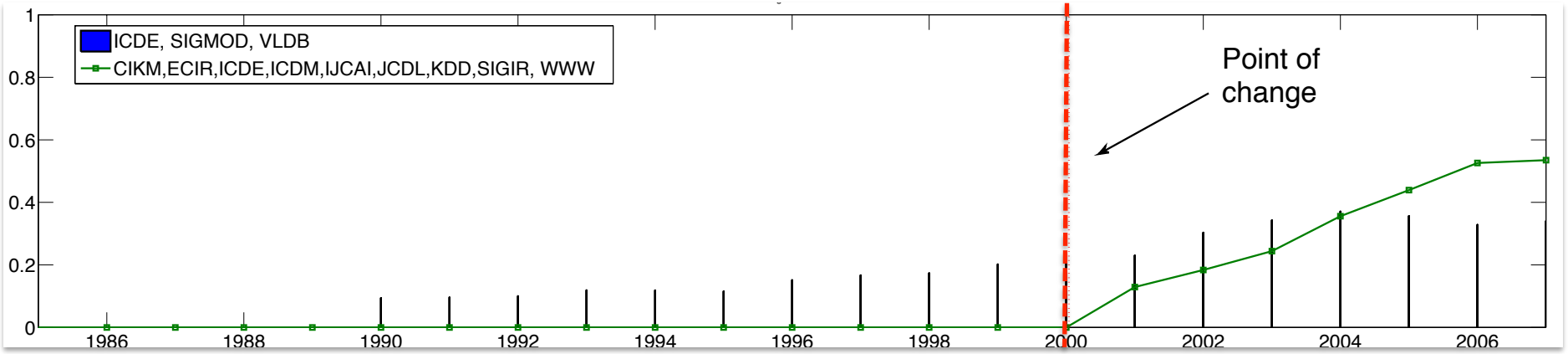
Outline

- Introduction to Tensors
- **Applications**
- Conclusions

Datasets

Name	Description	Dimensions
DBLP-1	(author, paper, conf)	$14.5K \times 14.4K \times 20$
DBLP-2	(author, conf, year)	$418K \times 3.5K \times 49$
LBNL	(src, dst, port #, time)	$65K \times 65K \times 65K \times 3.6K$
FACEBOOK	(wall, poster, day)	$64K \times 64K \times 1.8K$

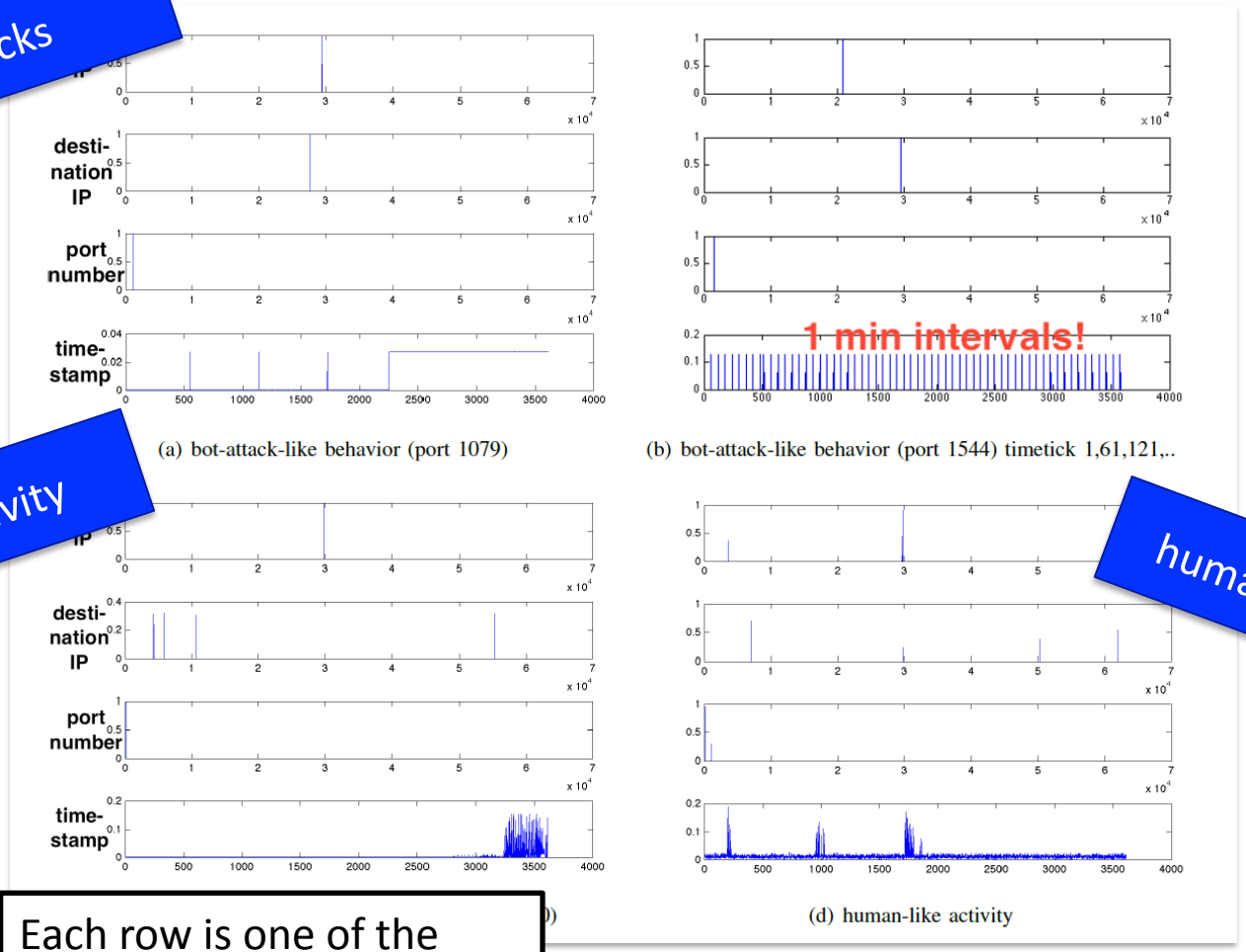
Application 1: Change Detection over Time



- 1st component: Database conferences
- 2nd component: Data mining venues
- Spotted known professor who changed area of research.

Application 2: Anomaly Detection in LBNL Network Traffic Data

bot-attacks



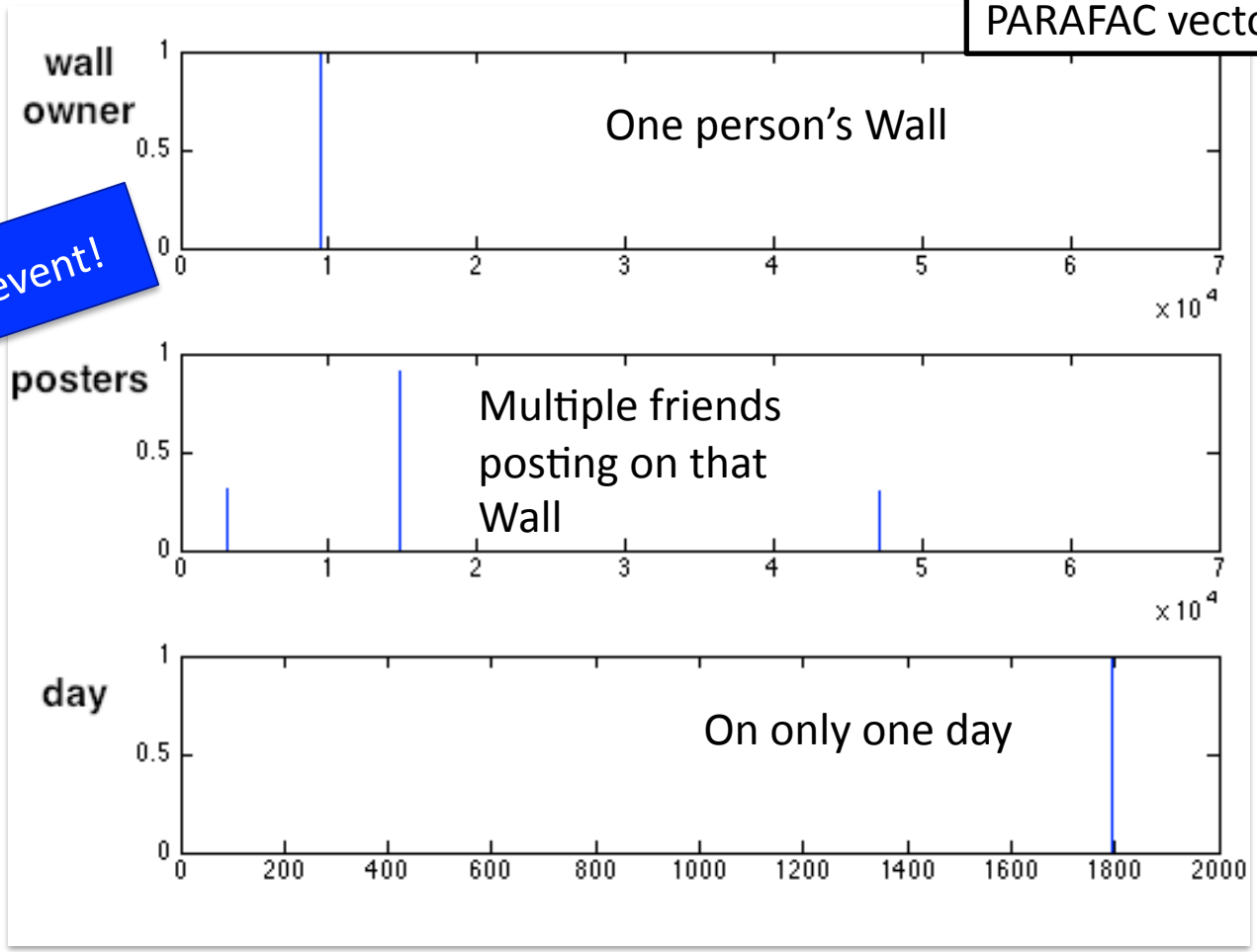
burst-of-activity

human-like activity

Each row is one of the PARAFAC vectors

Application 3: Anomaly Detection in Facebook

Each row is one of the PARAFAC vectors



Birthday-like event!

Application 3: Clustering in DBLP

Authors	Conferences
Bing Liu, Jure Leskovec, Christos Faloutsos, Hanghang Tong, Wynne Hsu	KDD
Andrew W. Moore, John Shawe-Taylor, Nello Cristianini, Michael I. Jordan	ICML
Michael J. Carey, H. V. Jagadish, Rakesh Agrawal, Divesh Srivastava, Christos Faloutsos	VLDB
Jeffrey F. Naughton, David J. DeWitt, Nancy E. Hal	SIGMOD
Vincent Conitzer, Tuomas Sandholm, Andrew Gilpin	AAAI

- clusters of authors publishing at the same venues
- advisor-advisee relationship between clustered authors
  - e.g. Christos Faloutsos and {Jure Leskovec, Hanghang Tong}

Outline

- Introduction to Tensors
- Applications
- **Conclusions**

Conclusions

- We propose a powerful way of modeling data that enables us to do:
  - Clustering
    - Clustering authors on the DBLP network
  - Anomaly Detection
    - Detecting network attacks and anomalies on Facebook
  - Change Detection in time
    - Detecting bridge authors on DBLP who gradually switch fields.

The End

Thank you!

For questions, please drop us an e-mail.

Special Tanks to Dimitra and Kostas!

Danai Koutra
Email: danai@cs.cmu.edu
Web: http://www.cs.cmu.edu/~dkoutra

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

Christos Faloutsos
Email: christos@cs.cmu.edu
Web: http://www.cs.cmu.edu/~christos