SAND2009-1268C

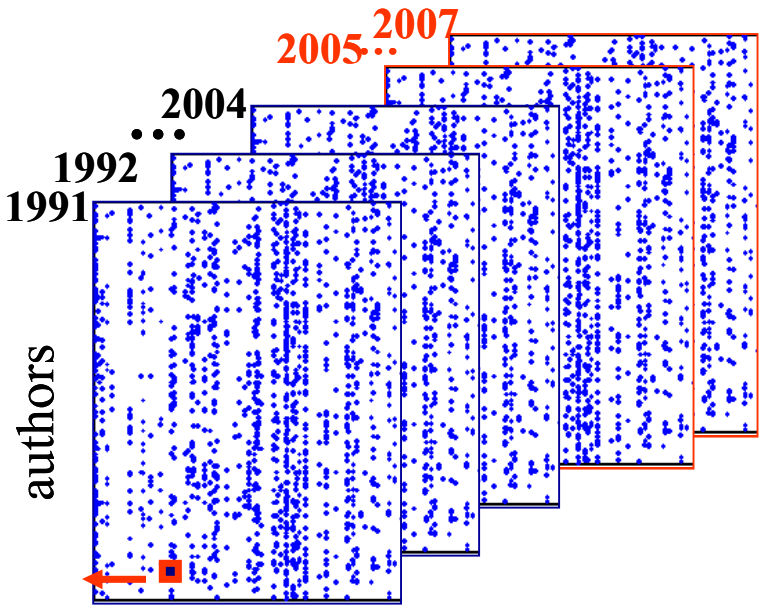
# Link Prediction on Evolving Data using Tensor Factorizations

**Evrim Acar, Tamara G. Kolda, Daniel M. Dunlavy**
Sandia National Labs., Livermore, CA


Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

# Problem Defn: Link Prediction

- Authors x Conferences x Years (10K x 2K x 14; ~0.1%)




$x_{ijk}$ = # of papers by $i^{th}$ author at $j^{th}$ conf. in year k.

Q1: Can we use tensor decompositions to model the data and extract meaningful underlying factors?

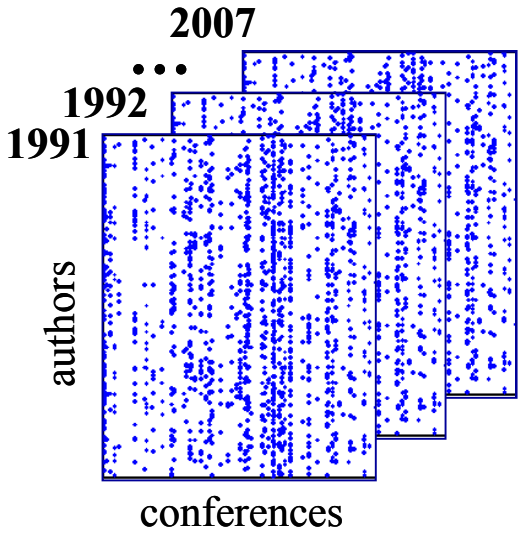
Q2: Can we predict who is going to publish at which conferences in future?

# Link Mining [Getoor and Diehl'05]

- **Not Interested in** missing link prediction [Clauset et al.'08]
- **Interested in** link prediction over time
  - Symmetric cases, e.g., co-authorship
    - Scores assigned to pairs based on network topology features [Liben-Novell and Kleinberg'07]
    - Binary classification problem [Al Hasan et al.'06]
  - Unsymmetric cases, e.g., who rated what
    - Binary classification problem [Liu and Kou'07]
    - Matrix Factorization Approaches [Koren'08]
    - Latent Space and Markov Models [Sarkar and Moore'05; Sarkar et al.'07]
- Datasets: arXiv, DBLP, BIOBASE, CiteSeer, NIPS, Netflix

- **Our Approach**
  - Tensor factorizations
    - CANDECOMP/PARAFAC (CP) model

      [Carroll & Chang'70; Harshman'70]
  - DBLP dataset (authors-conferences)

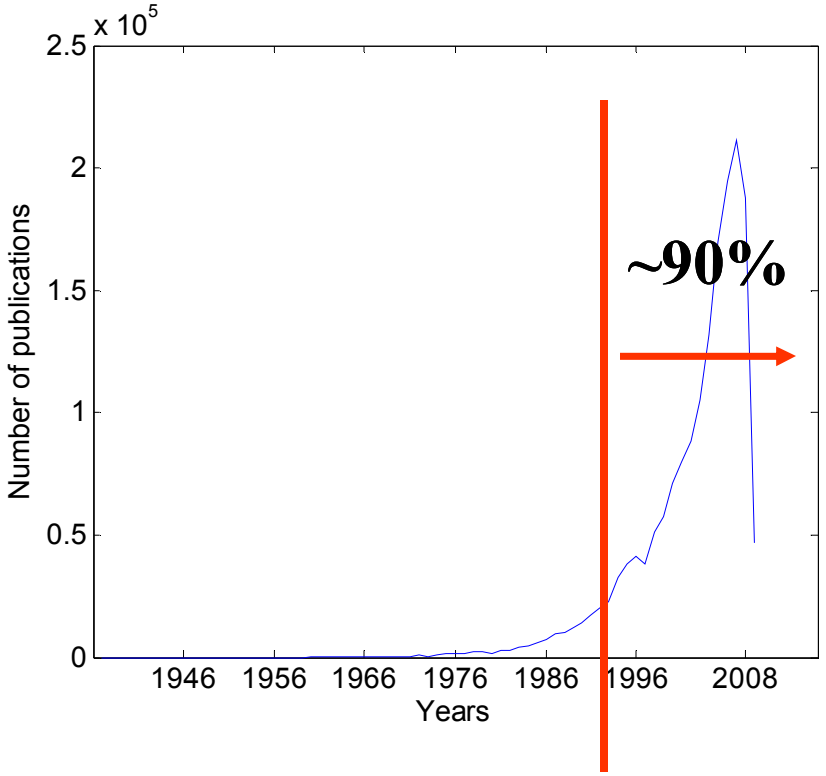


# Data Preprocessing

- **Available Data:**
  - **Publications:** ‘article’, ‘book’, ‘phdthesis’, ‘inproceedings’, etc.
  - **Time frame:** 1936-2008

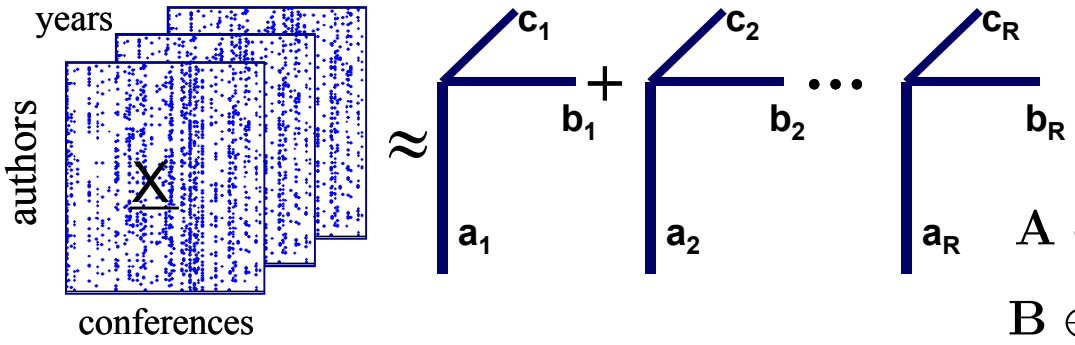
- **Our Data:**
  - **Publications:** ‘inproceedings’
  - **Time frame:** 1991-2007
    - Training Set: 1991 - 2004
    - Test Set: 2005 - 2007
  - **Authors:** more than 14 publications in the training period
  - Each entry:

$$x_{ijk} = log(x_{ijk}) + 1 \text{ if } x_{ijk} > 0$$



# Modeling DBLP using CP

$\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$




$$\mathbf{A} \in \mathbb{R}^{I \times R} = \left[\mathbf{a}_1 \cdots \mathbf{a}_R\right]$$
$$\mathbf{B} \in \mathbb{R}^{J \times R} = \left[\mathbf{b}_1 \cdots \mathbf{b}_R\right]$$
$$\mathbf{C} \in \mathbb{R}^{K \times R} = \left[\mathbf{c}_1 \cdots \mathbf{c}_R\right]$$

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \underline{\mathbf{X}} - \left( \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \right) \right\|^2$$

$$\|\underline{\mathbf{X}}\| = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{I} x_{ijk}^2$$

**Traditional approach:**
**Alternating Least Squares (ALS)**

**Alternative approach:**
**Gradient-based optimization, which is robust to overfactoring**
**[Acar, Kolda and Dunlavy'09]**

# Components make sense!

hans peter meinzer

Refine by AUTHOR

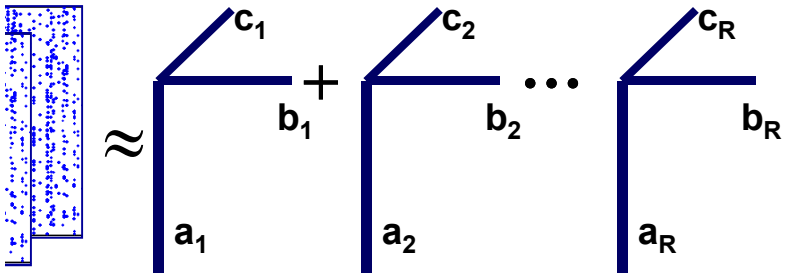
Hans-Peter Meinzer (136)
Ivo Wolf (40)
Matthias Thorn (24)
Gerald-P. Glombitza (24)
[top 4] [top 50] [all 187]

Refine by VENUE

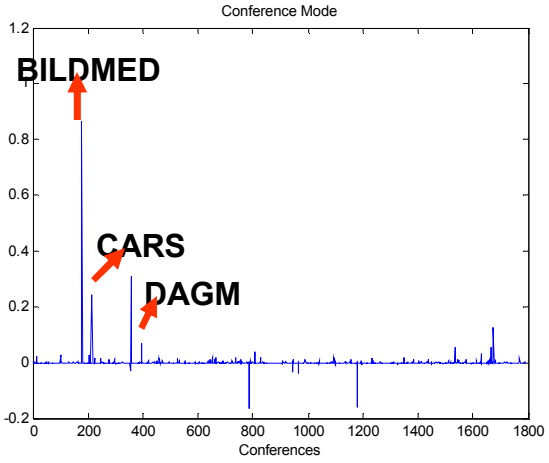
Bildverarbeitung für die Medizin (76)
DAGM-Symposium (12)
CARS (8)
Rechner- und sensorgestützte Chirurgie (6)
[top 4] [all 26]

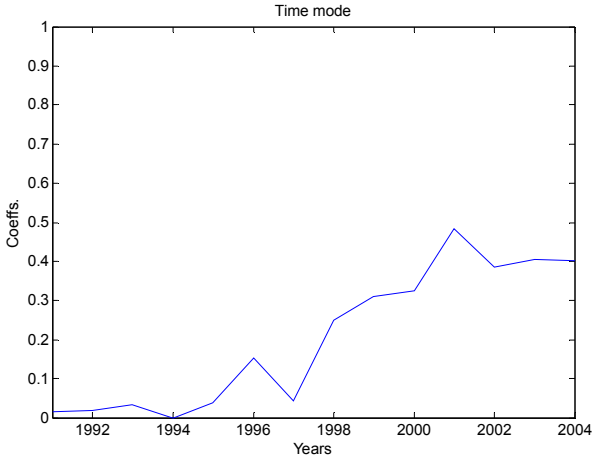
Refine by YEAR

2001 (17)
1998 (11)
2000 (11)
2004 (10)
2007 (10)
2008 (10)
1999 (10)
2003 (9)
2002 (8)
2006 (6)




$b_r$ $c_r$



# Components make sense!

Craig Boutilier

Refine by AUTHOR
Craig Boutilier (116)
Pascal Poupart (13)
Moisés Goldszmidt (9)
Ronen I. Brafman (8)
[top 4] [top 50] [all 66]

Refine by VENUE
UAI (25)
IJCAI (20)
AAAI (14)
AAAI/IAAI (12)
[top 4] [all 26]

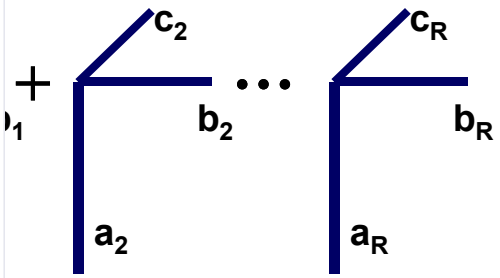
Refine by YEAR
2003 (10)
1999 (10)
1996 (8)
2000 (8)
[top 4] [all 20]

Daphne Koller

Refine by AUTHOR
Daphne Koller (157)
Joseph Y. Halpern (21)
Nir Friedman (17)
Eran Segal (14)
[top 4] [top 50] [all 118]

Refine by VENUE
UAI (33)
NIPS (19)
IJCAI (15)
AAAI/IAAI (13)
[top 4] [all 48]

Refine by YEAR
2003 (16)
2001 (13)
2000 (13)
2006 (11)
[top 4] [all 21]




21. IJCAI 2009: Pasadena, California, US

20. IJCAI 2007: Hyderabad, India

Manuela M. Veloso (Ed.): IJCAI 2007, Proceedings of the 20th Intern
Contents BibTeX

Thomas S. Huang, Anton Nijholt, Maja Pantic, Alex Pentland (Eds.): A
and Invited Papers. Lecture Notes in Computer Science 4451 Springer
Contents BibTeX

Artur S. d'Avila Garcez, Pascal Hitzler, Guglielmo Tamburrini (Eds.): P
230 CEUR-WS.org 2007
Contents BibTeX

19. IJCAI 2005: Edinburgh, Scotland, UK

Leslie Pack Kaelbling, Alessandro Saffiotti (Eds.): IJCAI-05, Proceedi
Contents BibTeX - IJCAI 2005 Home Page

18. IJCAI'2003: Acapulco, Mexico

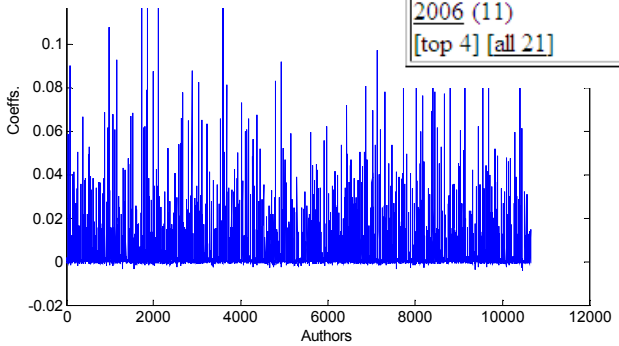
Georg Gottlob, Toby Walsh (Eds.): IJCAI-03, Proceedings of the Eigh
Contents BibTeX - IJCAI 2003 Home Page

Subbarao Kambhampati, Craig A. Knoblock (Eds.): Proceedings of IJ
Contents BibTeX - IIWeb 2003 Home Page

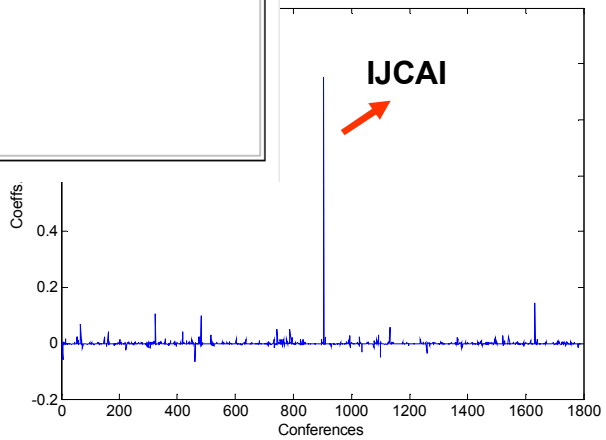
Bamshad Mobasher, Sarabjot S. Anand (Eds.): Intelligent Techniques f
ISBN 3-540-29846-0
Contents BibTeX

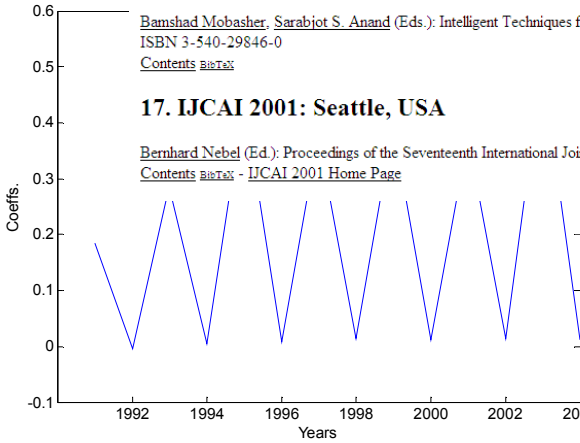
17. IJCAI 2001: Seattle, USA

Bernhard Nebel (Ed.): Proceedings of the Seventeenth International Joi
Contents BibTeX - IJCAI 2001 Home Page




nference mode

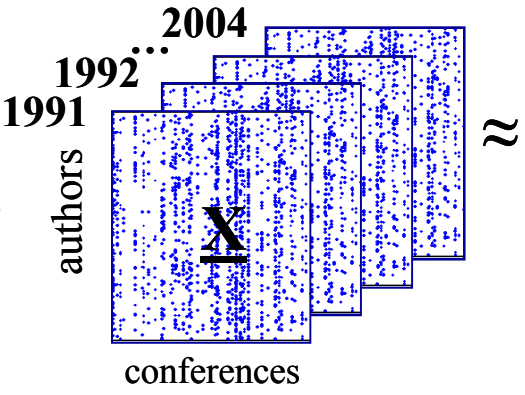


# Link Prediction Problem

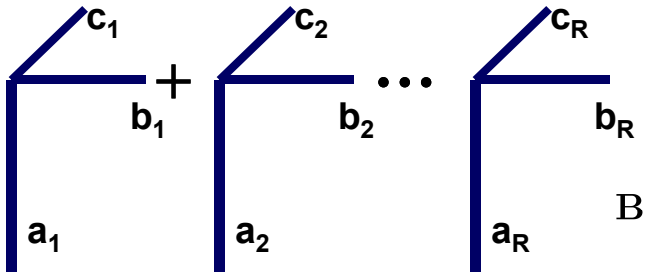
**TRAIN:**

$\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K_1}$




$\approx$

$\mathbf{C} \in \mathbb{R}^{K_1 \times R} = \left[\mathbf{c}_1 \cdots \mathbf{c}_R\right]$

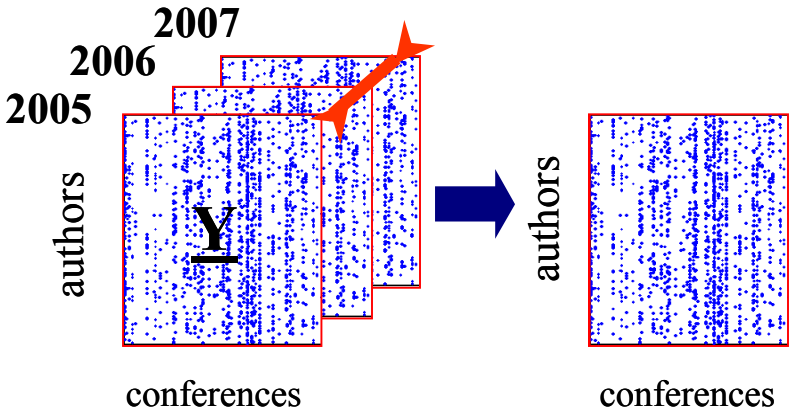



$\mathbf{B} \in \mathbb{R}^{J \times R} = \left[\mathbf{b}_1 \cdots \mathbf{b}_R\right]$

$\mathbf{A} \in \mathbb{R}^{I \times R} = \left[\mathbf{a}_1 \cdots \mathbf{a}_R\right]$

**TEST:**

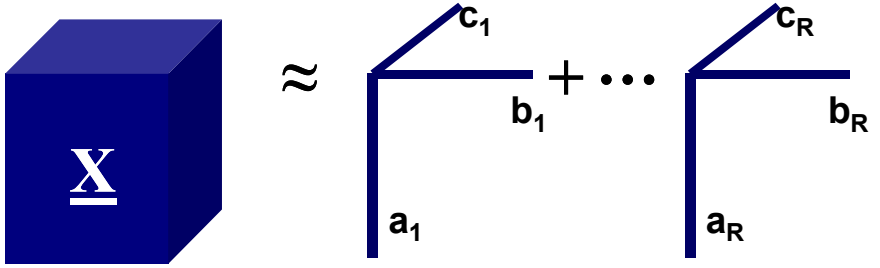
$\underline{\mathbf{Y}} \in \mathbb{R}^{I \times J \times K_2}$




~ 0.3% dense
~ 60K links (~ 32K new)

# Score for <author$_i$, conf$_j$>

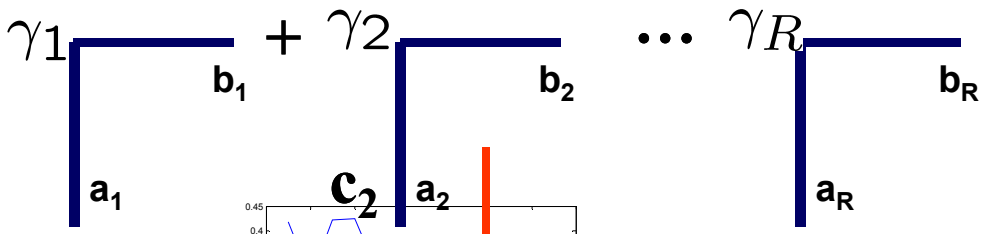
- Sign ambiguity:

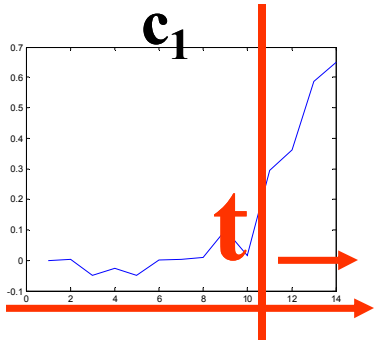


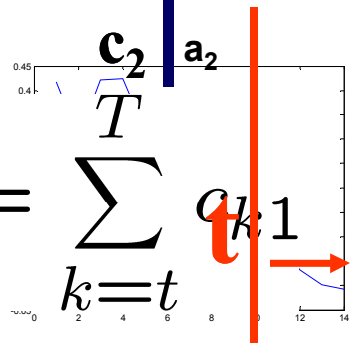

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{2} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$
$$\approx (-\mathbf{a}_1) \circ (-\mathbf{b}_1) \circ \mathbf{c}_1 + \mathbf{a}_2 \circ (-\mathbf{b}_2) \circ (-\mathbf{c}_2)$$
$$\approx (-\mathbf{a}_1) \circ \mathbf{b}_1 \circ (-\mathbf{c}_1) + (-\mathbf{a}_2) \circ (-\mathbf{b}_2) \circ \mathbf{c}_2$$

- Fix signs using the signs of the maximum magnitude entries and then compute a score for each author-conference pair using the information from time domain:

$$\mathbf{SCORE}_{ij} = |\sum_{r=1}^{R} \gamma_r \mathbf{a}_{ir} \mathbf{b}_{jr}|$$



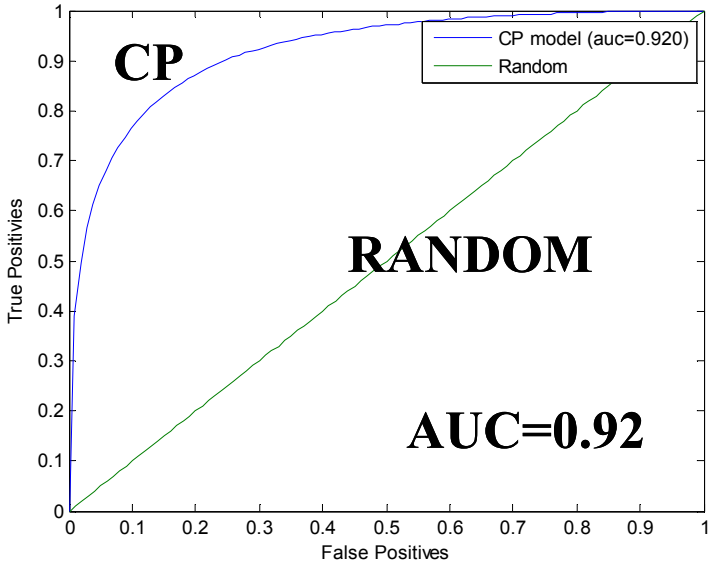



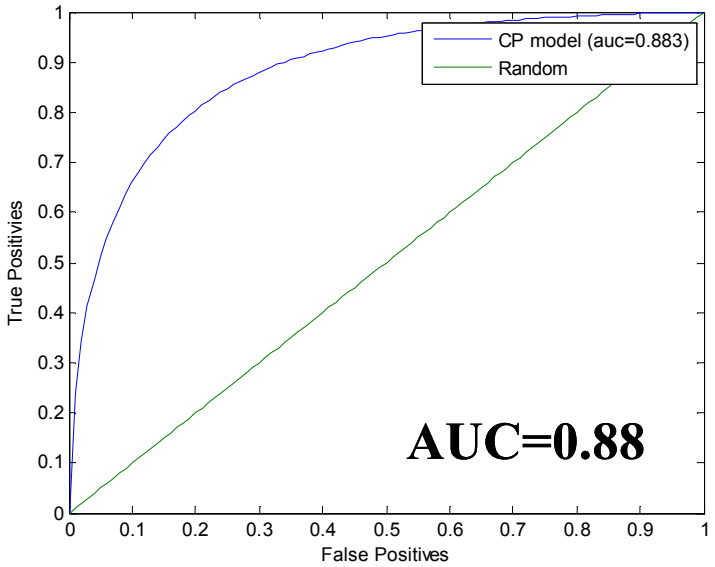

$$\gamma_1 = \sum_{k=t}^{T} c_{k1} \qquad \gamma_2 = \sum_{k=t}^{T} c_{k2}$$

# Performance Evaluation

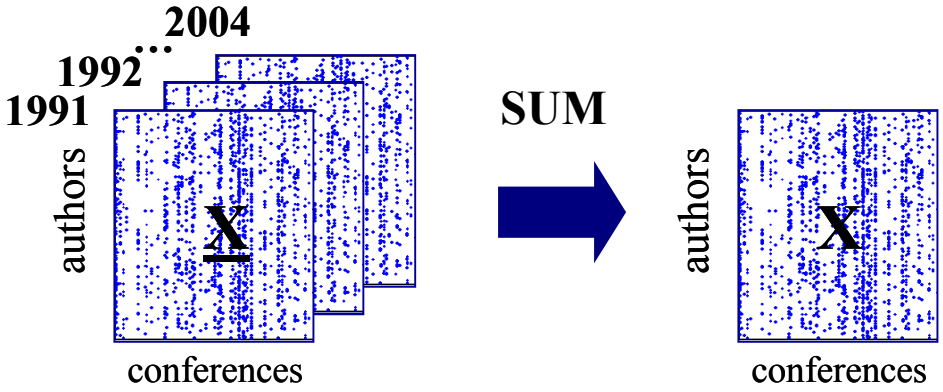
Predicting **All Links** between 2005 and 2007 (~ 60K):



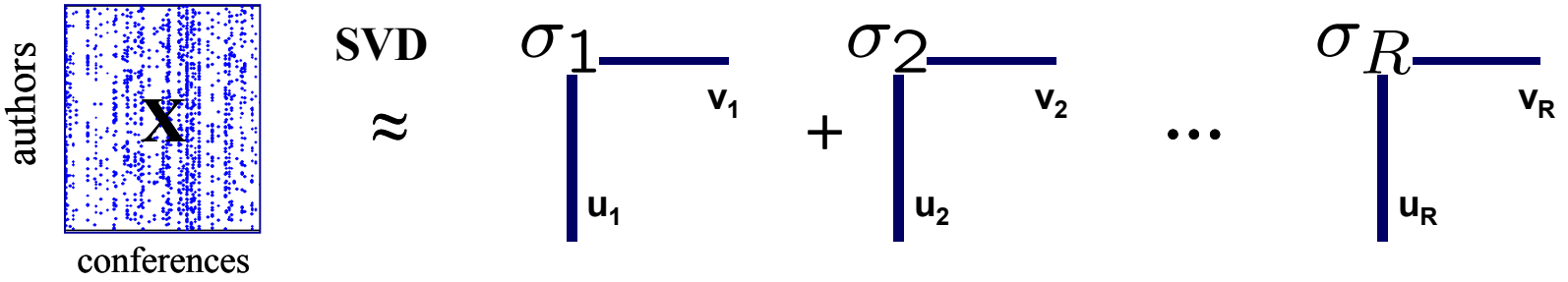

Predicting **Only New Links** between 2005 and 2007(~ 32K):



# Alternative Approach: SVD



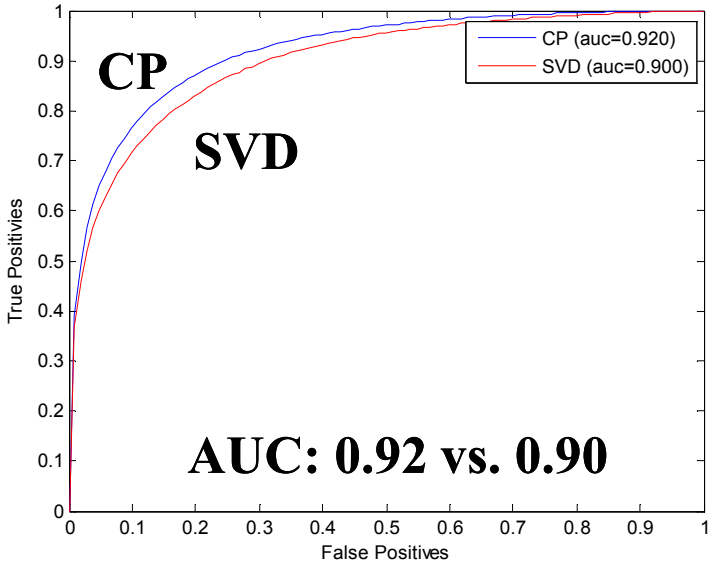

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$$



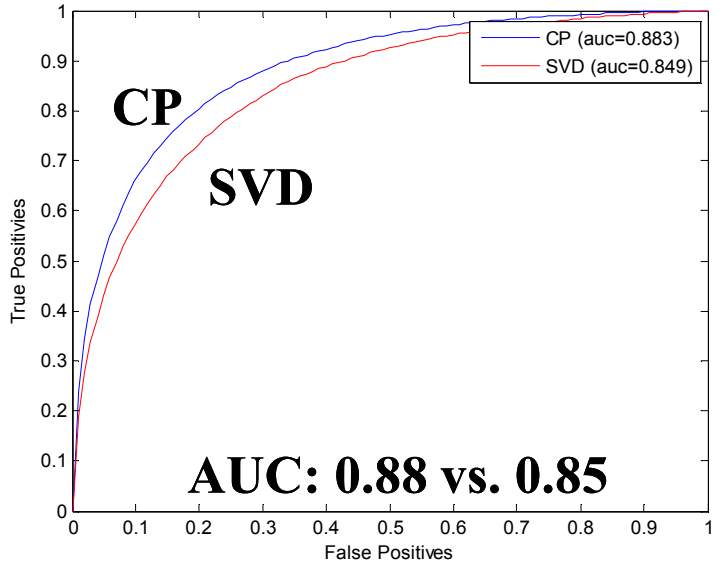

$$\mathbf{SCORE}_{ij} = |\sum_{r=1}^{R} \sigma_r \mathbf{a}_{ir} \mathbf{b}_{jr}|$$

# CP vs. SVD

Predicting **All Links** between 2005 and 2007 (~ 60K):

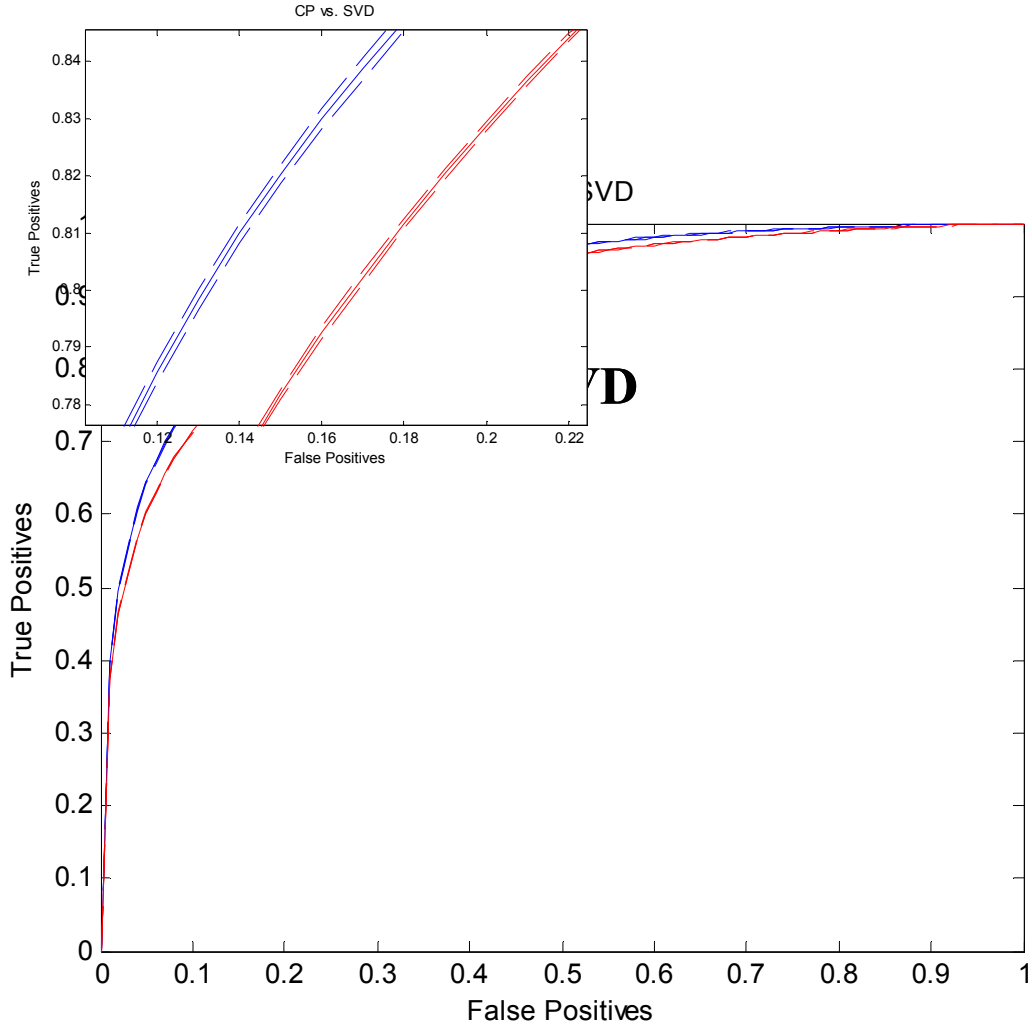



Predicting **Only New Links** between 2005 and 2007(~ 32K):



# Sensitivity Analysis

Leave-out 10% of the data and model the rest:



# More work in progress

- **Algorithm:**
  - **Initialization:** First two modes using HOSVD ($I, J \leq R$); last mode **RANDOM** ($K>R$)
- **Model:**
  - **Number of components (*R=50*):** Core consistency diagnostic, overall model fit, how well the model fits to nonzeros.
  - **Time Domain info:** better ways of using the time domain information
  - **Sparsity constraints:**

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \underline{\mathbf{X}} - \left( \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \right) \right\|^2 + ???$$

- **Performance Evaluation:**
  - **Statistical Properties of the model:** Can we make use of jackknifing/bootstrapping?
  - **Evaluation metrics:** Metrics that could be more informative than AUC

# Thank you!

- **References:**
  - Acar, Kolda and Dunlavy, *An Optimization Approach for Fitting Canonical Tensor Decompositions*, SAND2009-0857, Feb. 2009.
  - Toolboxes by Sandia:
    - Tensor Toolbox
    - Poblano Toolbox
  - The DBLP Computer Science Bibliography, http://www.informatik.uni-trier.de/~ley/db/
- **Contact:**
  - Evrim Acar, **eacarat@sandia.gov**
  - Tamara G. Kolda, **tgkolda@sandia.gov**
  - Daniel M. Dunlavy, **dmdunla@sandia.gov**