LogPar: Logistic PARAFAC2 Factorization for Temporal Binary Data with Missing Values

Kejing Yin, Ardavan Afshar, Joyce C. Ho, William K. Cheung, Chao Zhang, Jimeng Sun

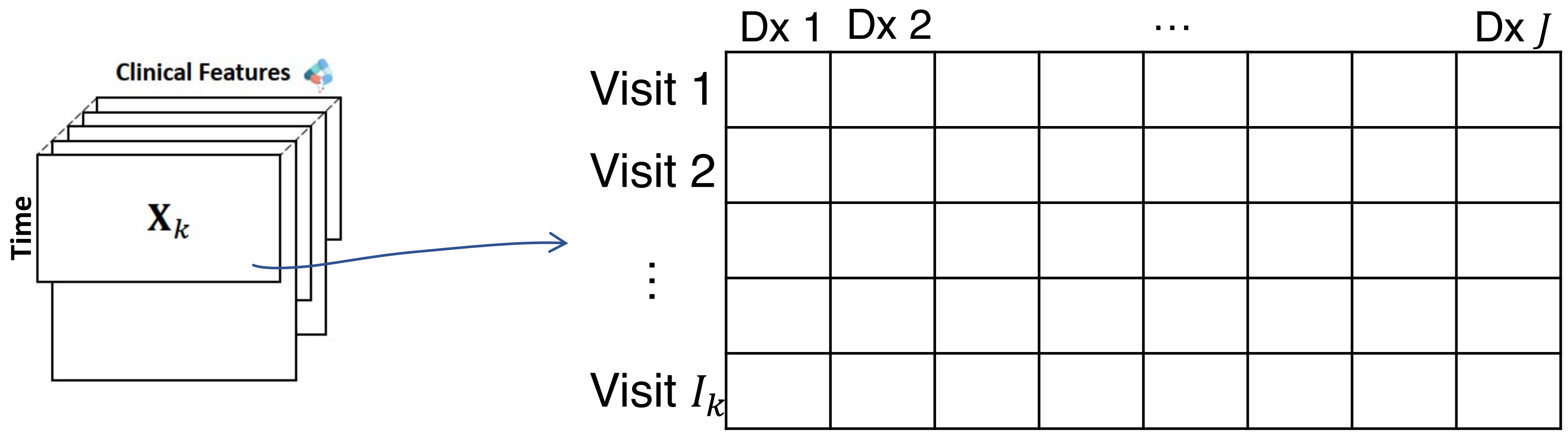
Department of Computer Science, Hong Kong Baptist University
SUNLAB
Georgia Tech
Emory University

Background: Binary Irregular Tensors

- Real-world data are often collected over time.

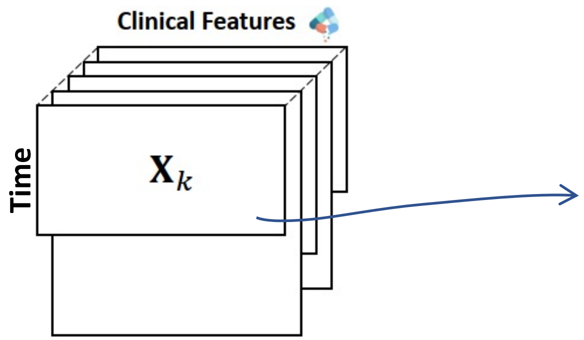
- Irregular Tensor: A collection of matrices with varying sizes in one dimension.
- Handles the temporally irregular data.

- Example: Electronic health records



- Other applications: spatio-temporal modeling, recommender systems, etc.

Background: Implicit one-class missingness



	Dx 1	Dx 2			…			Dx J
Visit 1	1	0	0	0	0	0	0	0
Visit 2	0	0	1	0	0	0	1	0
	1	0	0	0	0	0	0	0
⋮	0	0	0	1	0	0	0	0
Visit I_k	0	1	0	0	0	0	0	0

- Values of 1s: confirmed diagnosis codes assigned by doctors.
- Values of 0s?
  - The patient does have the disease.
  - The diagnosis was not performed: **missing data**.

- We do not know the missing entries (**implicit**) & we only consider missing 1s (**one-class**)

Research Questions

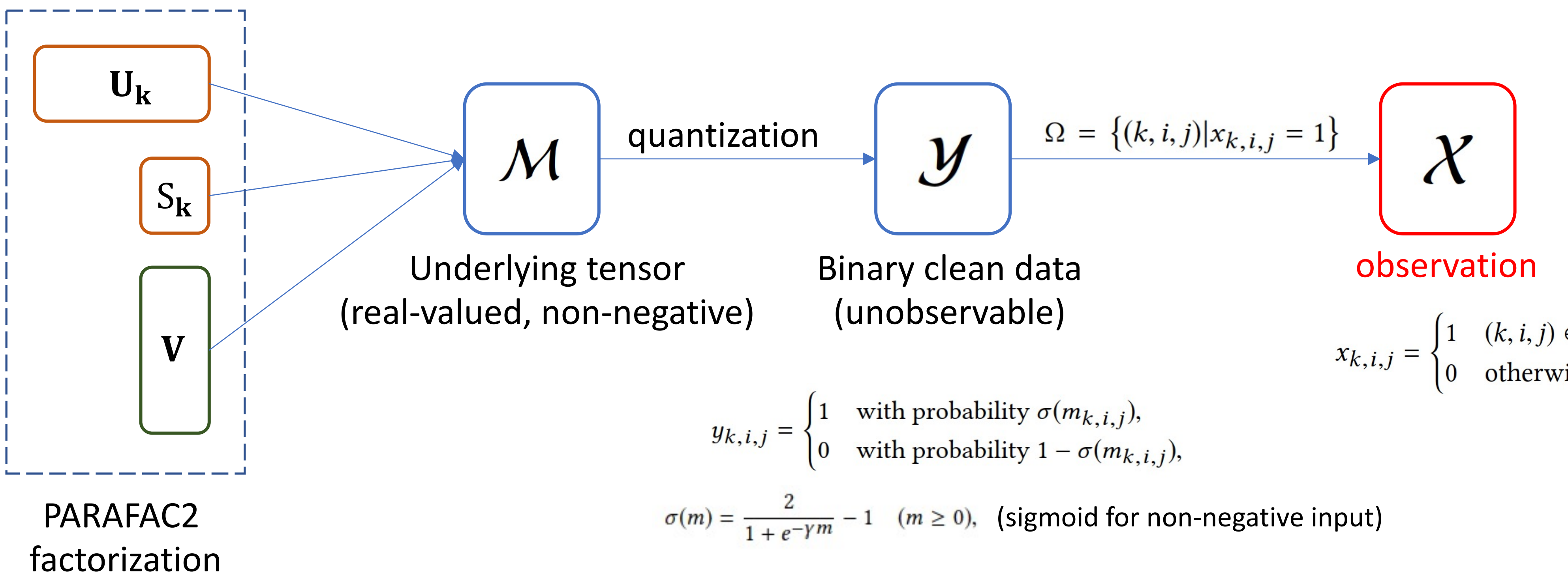
- How to complete the missing data in binary irregular tensors?
- How to discover meaningful latent factors from such data?

Our solution: Logistic PARAFAC2 (LogPar)

Problem:

Given a temporal binary irregular tensor $\mathcal{X}$, discover its latent factors and completion.

Observation model for binary data:



$$x_{k,i,j} = \begin{cases} 1 & (k,i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

$$y_{k,i,j} = \begin{cases} 1 & \text{with probability } \sigma(m_{k,i,j}), \\ 0 & \text{with probability } 1 - \sigma(m_{k,i,j}), \end{cases}$$

$$\sigma(m) = \frac{2}{1+e^{-\gamma m}} - 1 \quad (m \geq 0),$$ (sigmoid for non-negative input)

Our solution: Logistic PARAFAC2 (LogPar)

Handling the implicit one-class missing

Extending nnPU learning to logistic PARAFAC2:

$$\widetilde{R}_{\mathrm{pu}}(g) = \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) + \max\left\{0, \widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g)\right\}$$

$$\widehat{R}_{\mathrm{p}}^{+}(g) = (1/n_{\mathrm{p}}) \textstyle\sum_{i=1}^{n_{\mathrm{p}}} \boxed{\ell(g(x_i^{\mathrm{p}}), +1)}$$
$$\widehat{R}_{\mathrm{p}}^{-}(g) = (1/n_{\mathrm{p}}) \textstyle\sum_{i=1}^{n_{\mathrm{p}}} \boxed{\ell(g(x_i^{\mathrm{p}}), -1)}$$
$$\widehat{R}_{\mathrm{n}}^{-}(g) = (1/n_{\mathrm{n}}) \textstyle\sum_{i=1}^{n_{\mathrm{n}}} \boxed{\ell(g(x_i^{\mathrm{n}}), -1)}$$

Sample-wise loss

- Replace the loss function with the point-wise loss of logistic PARAFAC2

$$\ell(\widehat{x}_{k,i,j}, x_{k,i,j}) = \left(x_{k,i,j} \log \widehat{x}_{k,i,j} + (1 - x_{k,i,j}) \log(1 - \widehat{x}_{k,i,j})\right)$$

- Leading to the final objective function: $\tilde{\mathcal{L}}\left(\widehat{\mathcal{X}}\right) = \pi \frac{\langle \mathcal{X},\ \log(\widehat{\mathcal{X}})\rangle}{\|\mathcal{X}\|_1} + \max\left\{0, \frac{\langle 1-\mathcal{X},\ \log(1-\widehat{\mathcal{X}})\rangle}{\|1-\mathcal{X}\|_1} - \pi \frac{\langle \mathcal{X},\ \log(1-\widehat{\mathcal{X}})\rangle}{\|\mathcal{X}\|_1}\right\}$

Kiryo, Ryuichi, et al. "Positive-unlabeled learning with non-negative risk estimator." *Advances in Neural Information Processing Systems*. 2017.

Our solution: Logistic PARAFAC2 (LogPar)

Regularizations for Better Interpretability

- Uniqueness Regularization:

$$\mathcal{R}_1 = \sum_{k=1}^{K} \frac{\mu}{2} \left\| \mathbf{U}_k{}^{\top} \mathbf{U}_k - \boldsymbol{\Phi} \right\|_F^2$$

(promotes factor invariance)

- Time-Aware Temporal Smoothing:

$$\mathcal{R}_2 = \sum_{k=1}^{K} \sum_{i=2}^{I_k} e^{-\beta \delta_i} \left| \mathbf{u}_{k,t} - \mathbf{u}_{k,t-1} \right|$$

(discovers smoother temporal factors)

Experiments and Results

- **Datasets**
  - **Sutter**: collected from Sutter health, a large US based health provider network. We use diagnoses and medications of each clinical visit of patients.
  - **CMS**: publicly available CMS Linkable 2008-2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF). We use diagnoses of each clinical visit of patients.
  - **MIMIC-III**: a large-scale ICU dataset. We use medications and abnormal lab tests.

	Sutter	CMS	MIMIC-III
#Patients (K)	34,905	74,153	28,485
#Features (J)	328	319	405
Median(I_k)	26	26	22
Average(I_k)	30.5	29	28.4
#Positive entries	2.3M	4.5M	14.5M
Sparsity	0.80%	0.65%	4.43%
Single-feature visits	29.5%	37.4%	0.67%
Predictive task	Heart failure	–	Mortality
Positive label ratio	8.92%	–	8.86%

- **Baselines**
  - **COPA**: SOTA PARAFAC2 factorization model.
  - **SPARTan**: PARAFAC2 for sparse data.
  - **PU-MC**: Binary matrix completion model based on PU learning.
  - **One-class MF (OCMF)**: SOTA binary matrix completion model based on sampling.

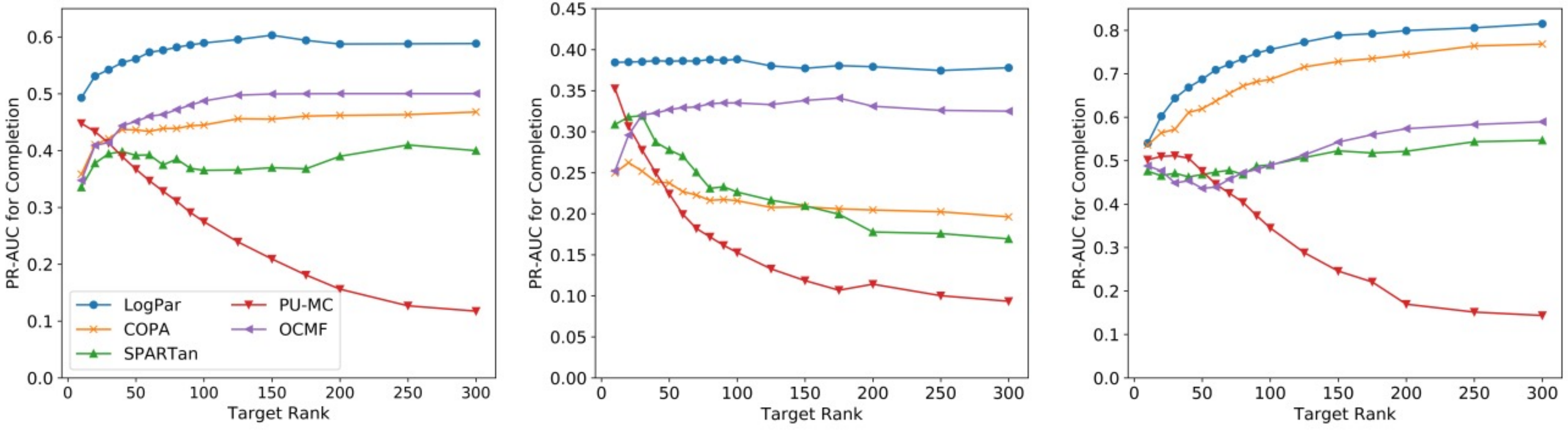
Experiments and Results

- **Experiment Setting**
  - Extract 10% positive entries for validation and parameter tuning.
  - Hold out 20% positive entries for test.
  - Randomly match 10 negative entries (value 0s) for each positive test entry as test set.
  - Use the remaining 70% positive entries as training set.

- **Evaluation Metric**
  - **PR-AUC**: binary & imbalanced

Experiments and Results

- **Completion performance with varying ranks**

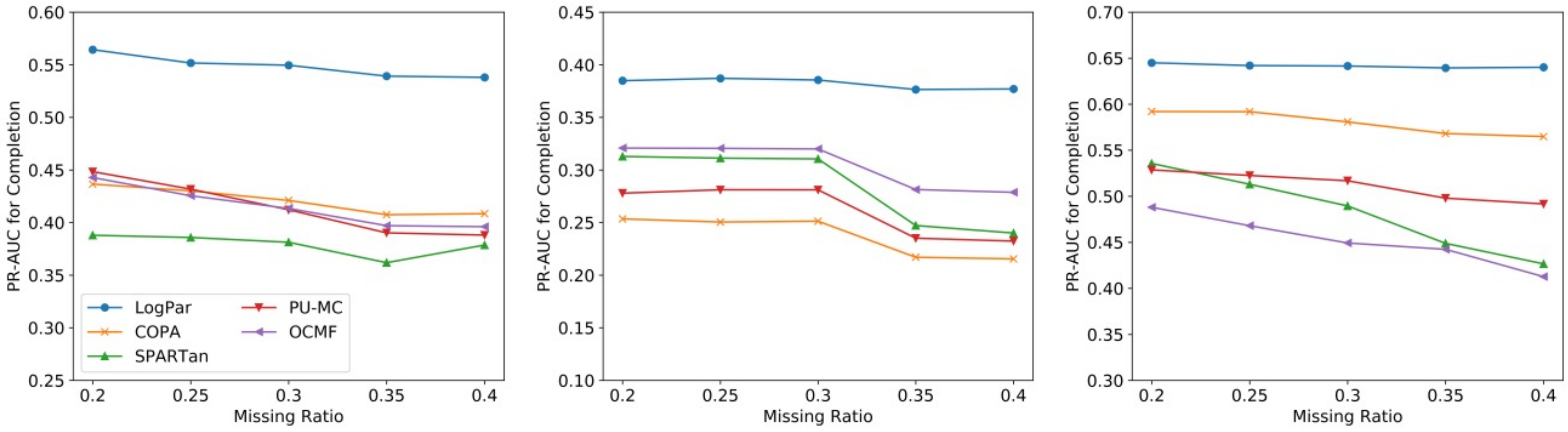


(a) Sutter Dataset (b) CMS Dataset (c) MIMIC-III Dataset

- LogPar consistently outperforms all baselines for all target ranks
- Performance of PU-MC decrease with increasing rank: severe overfitting.

Experiments and Results

- **Completion performance with varying missing ratios**
  - Fix rank of all models to 30

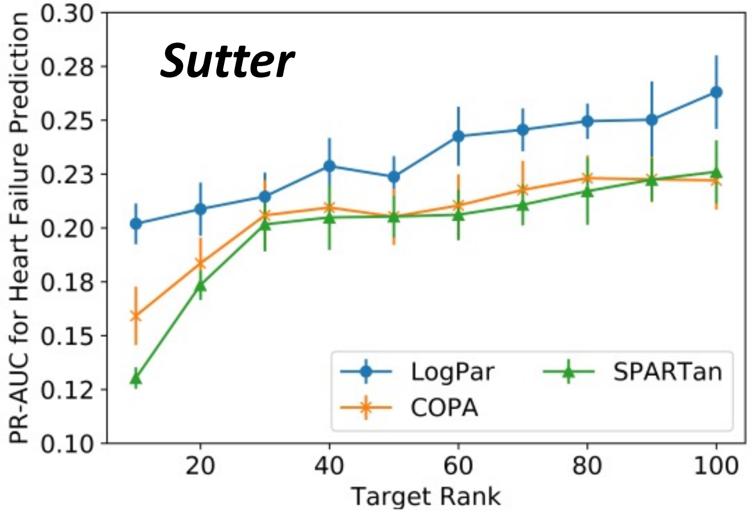


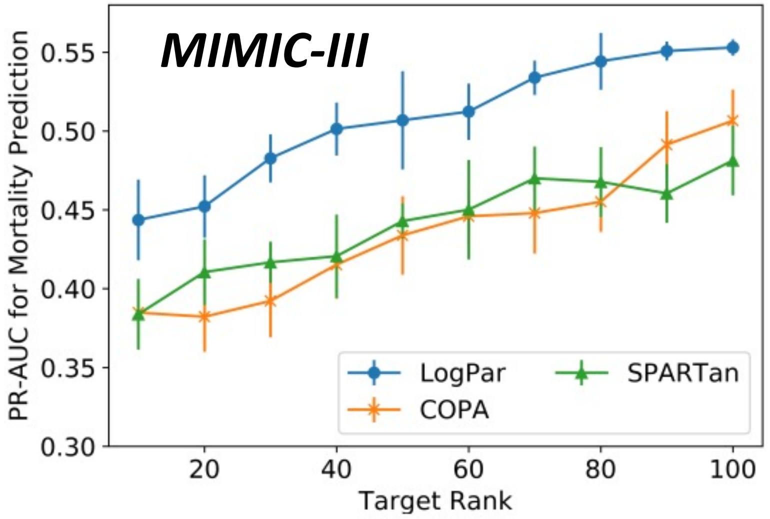
(a) Sutter Dataset (b) CMS Dataset (c) MIMIC-III Dataset

  - LogPar consistently outperforms all baselines for all level of missingness.

Experiments and Results

- **Downstream prediction tasks**
  - Fix missing ratio to 30% and split patients into training set and test set
  - Train LogPar with training set to learn the latent factor $\mathbf{V}$
  - Project test set onto the learned factor $\mathbf{V}$
  - Use $\boldsymbol{S}_k$ as features and train a logistic regression for classification. We use five-fold cross validation.





- **LogPar outperforms all baselines**

Conclusion

- We propose LogPar for binary irregular tensor factorization with explicit consideration of one-class missingness.
- LogPar is the **first PARAFAC2 model considering missingness**.
- We **incorporated the PU learning**, which greatly helps the completion.
- We **propose the uniqueness and the temporal smoothness regularization**.
- Empirical evaluation validates the effectiveness of LogPar and its components.

LogPar: Logistic PARAFAC2 Factorization for Temporal Binary Data with Missing Values

Kejing Yin, Ardavan Afshar, Joyce C. Ho, William K. Cheung, Chao Zhang, Jimeng Sun

Department of Computer Science, Hong Kong Baptist University
SUNLAB
Georgia Tech
Emory University