PERCeIDs: Periodic Community Detection

Lin Zhang, Alexander Gorovits, Petko Bogdanov

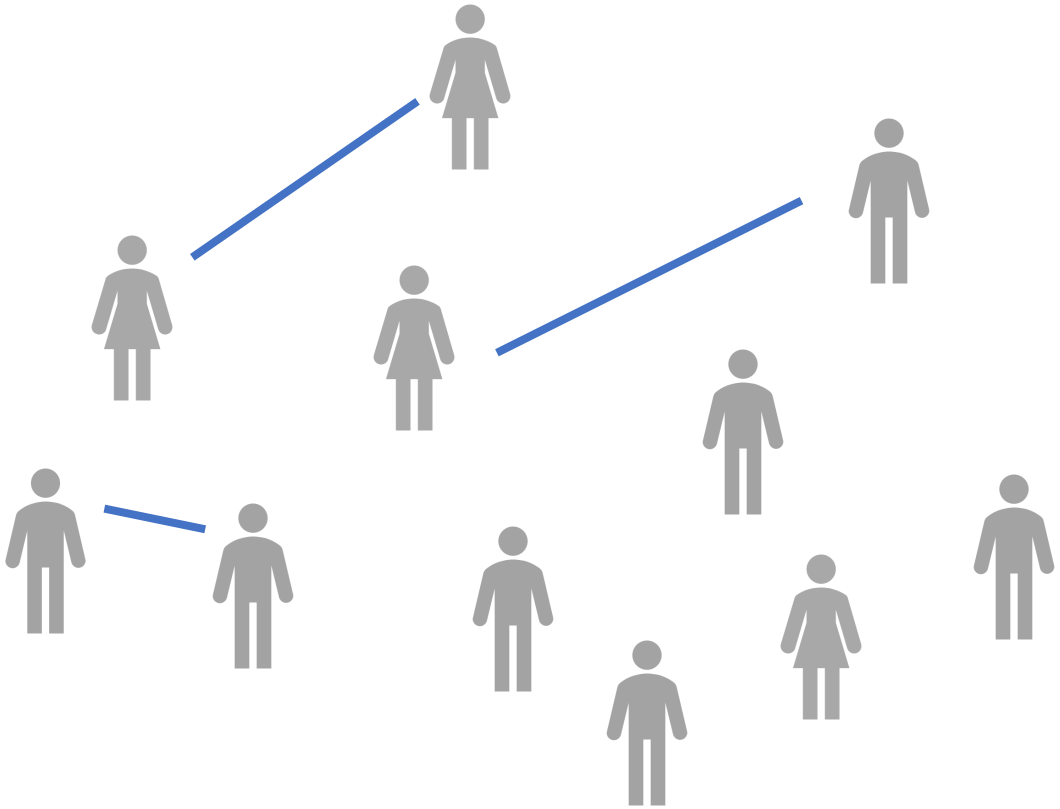
Department Of Computer Science, University at Albany-SUNY

Lin Zhang

Alex Gorovits

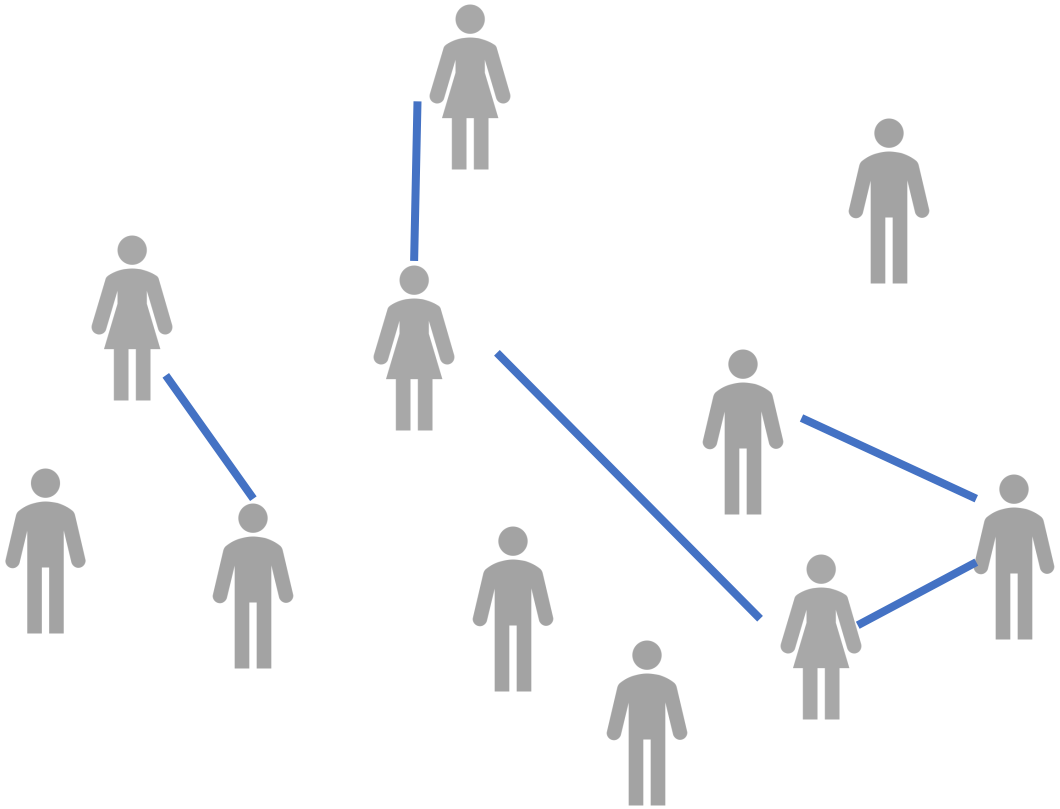
Temporal interactions

Day 1



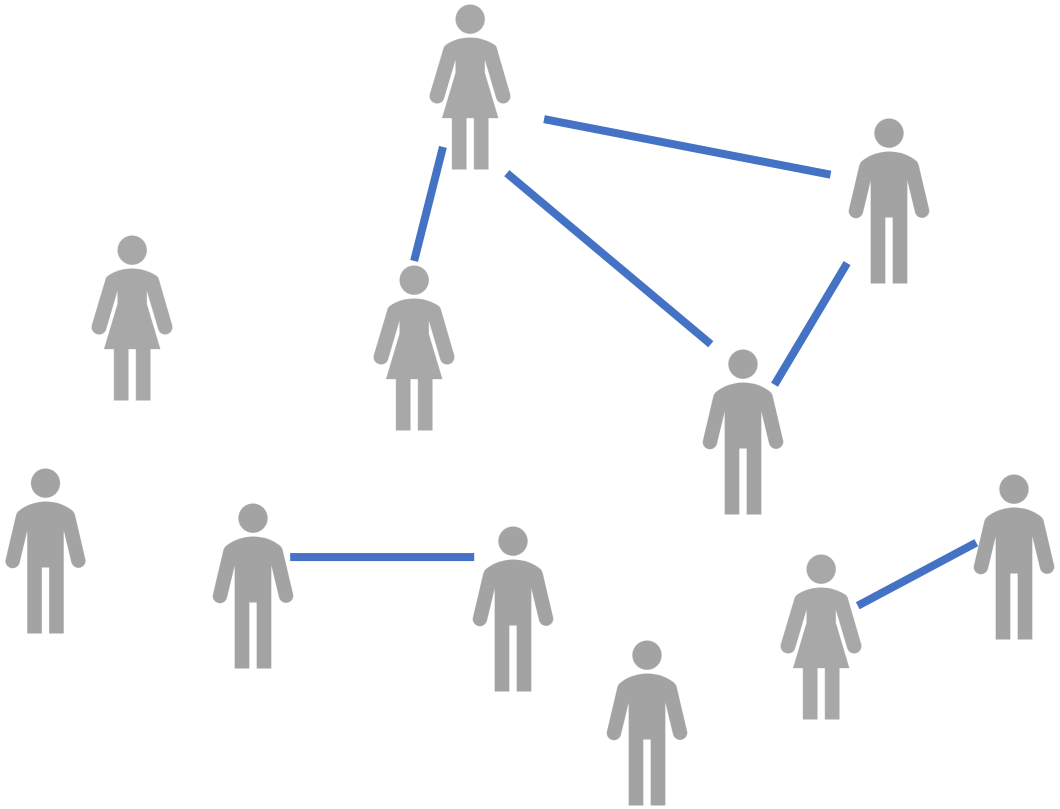
Temporal interactions

Day 2

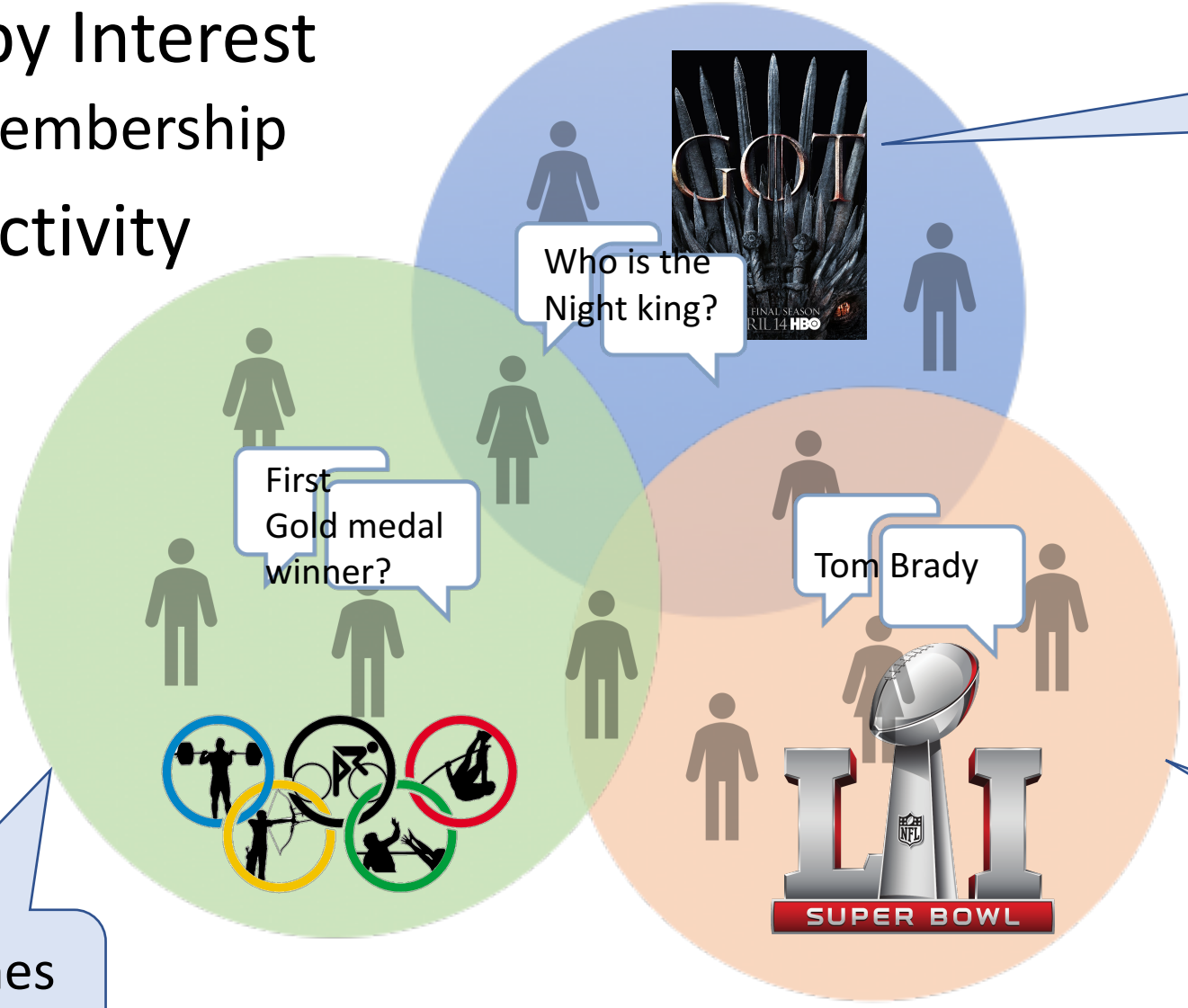


Temporal interactions

Day 3



- Grouped by Interest
  - Stable membership
- Periodic Activity

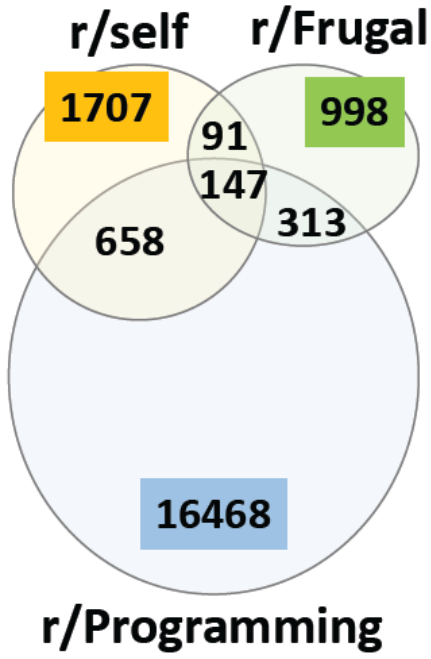


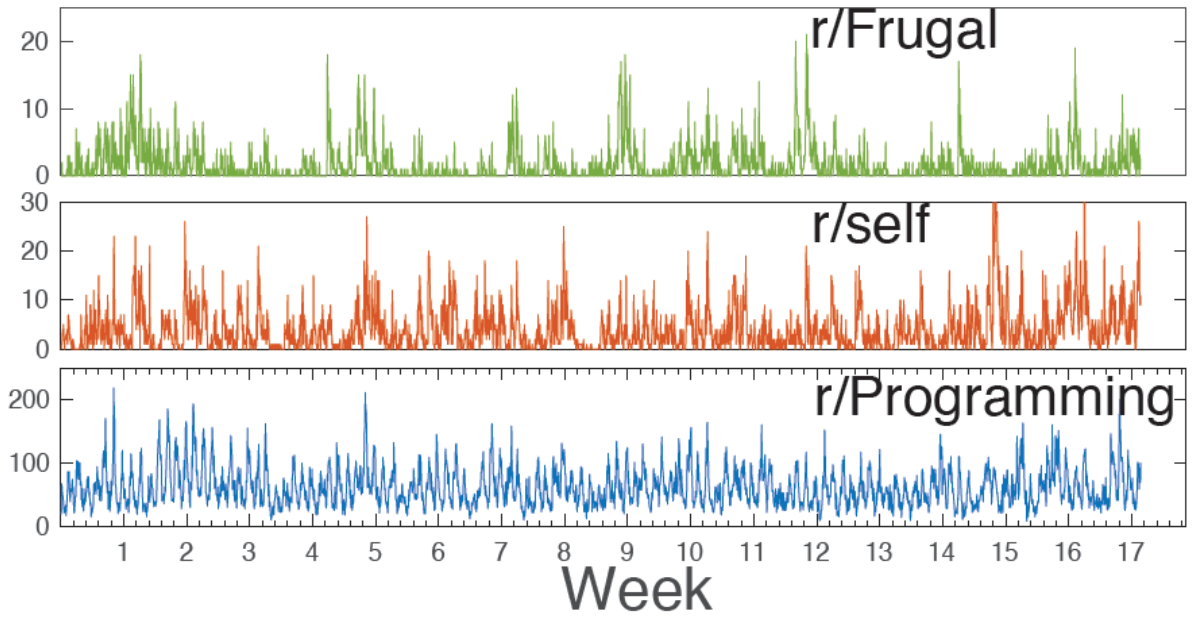
Game of Thrones (**weekly + yearly**)

Olympic games (**every 4 years**)

Super bowl (**yearly**)

Behavioral Periodic Communities

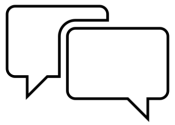




Significant member overlap, but distinct temporal activity

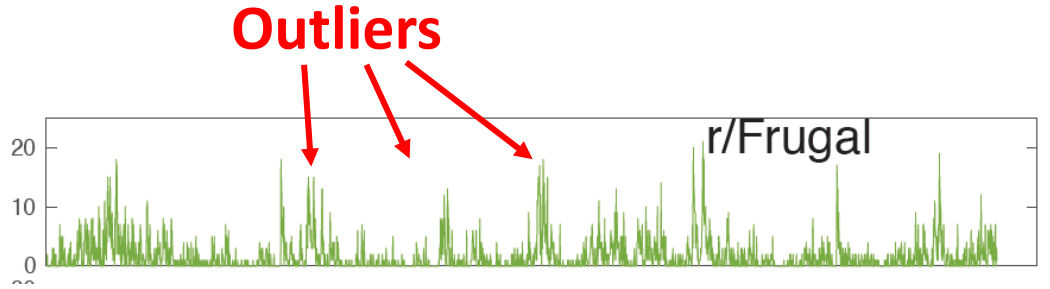
Applications

- Social networks
  - Context prediction: “advertise when the community is active”
  - Abnormal activity detection: “significant events that change activity pattern”
- Transportation networks
  - Optimal management of resources: public buses, subway stations, bike-sharing stations.
- Other networks:
  - Optimize the location of base stations of 4G, 5G, or WIFI.

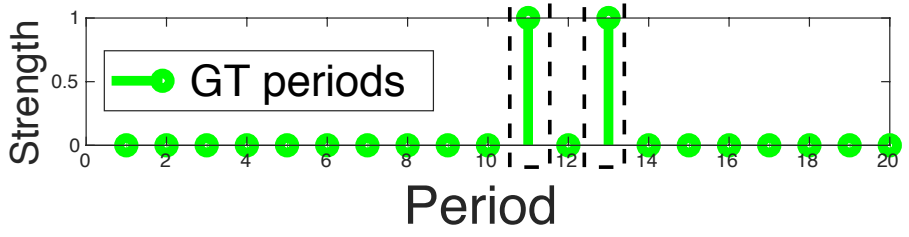


Challenges in periodic community detection

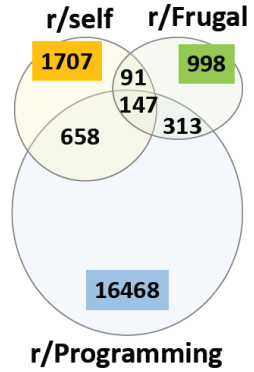
- Imperfect periodicity
  - **Outliers:** Abnormal events, such as ***Black Friday*** in the USA or ***11.11*** in China
  - **Noise:** Imperfections due to data collection
- Multiplicity of periods
  - Mixture of multiple periods, e.g. daily + weekly
- Overlapping community structure
  - A member may belong to multiple communities.



Multiple (mixture of) periods



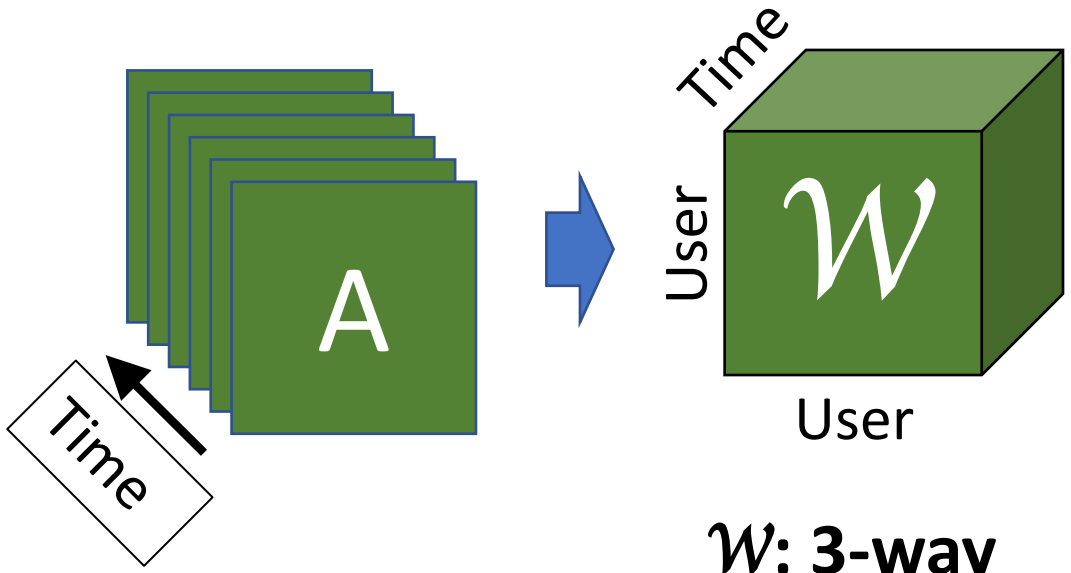
Overlapping membership



Related Work

- Communities in static networks
  - A variety of approaches (see *[Fortunato 2016]*, *[Xie 2013]*)
  - *[Yang 2013]* – CLAM, generative model for overlapping communities
    - No treatment/utilization of time
- Temporal Communities
  - Community/subgraph detection on dynamic graphs (*[Liu 2014]*, *[Mongiovi 2013]*, *[Ahmed 2011], many others*
    - Single subgraph per interval, often non-overlapping, sensitive to temporal aggregation
  - User-defined consistency *[Rozenshtein 2014]* or persistence (*[Liu 2014]*, *[Gorovits 2018]*)
    - "Simpler" on/off temporal structure, no model of seasonality

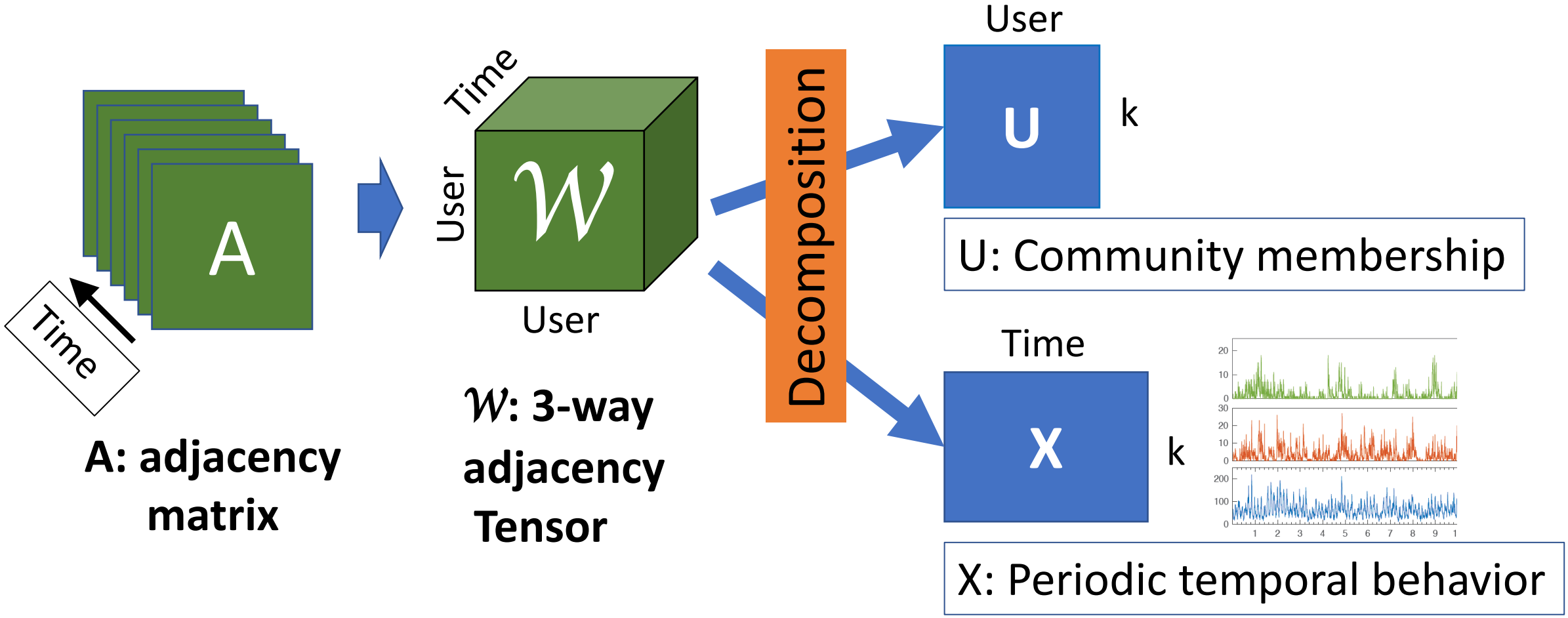
Model input as a tensor



A: adjacency matrix

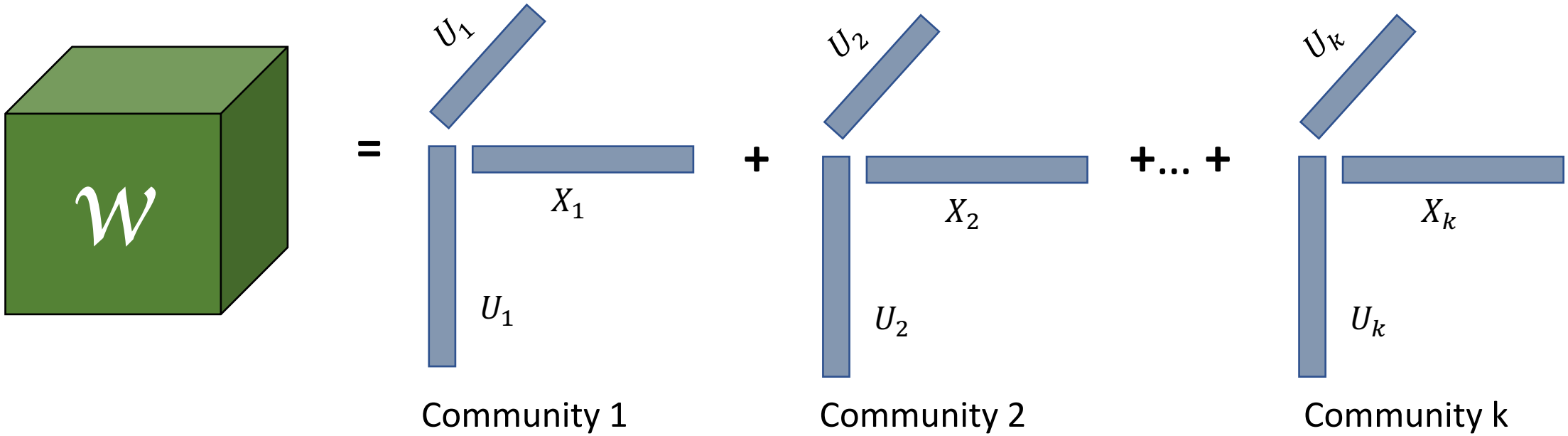
$\mathcal{W}$: 3-way adjacency Tensor

Periodicity-aware tensor decomposition



Data fitness: CP Tensor Factorization

- 3-way Tensor! $N_U \times N_U \times T$
- CANDECOMP/PARAFAC to decompose $\boldsymbol{W} \approx [[U, U', X]]$



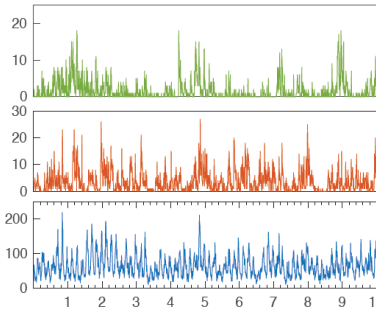
$$\mathrm{L}(W, U, X, k) = \left|\left|\mathcal{W} - [[U, U', X]]\right|\right|_F^2$$

Temporal Structure

Time

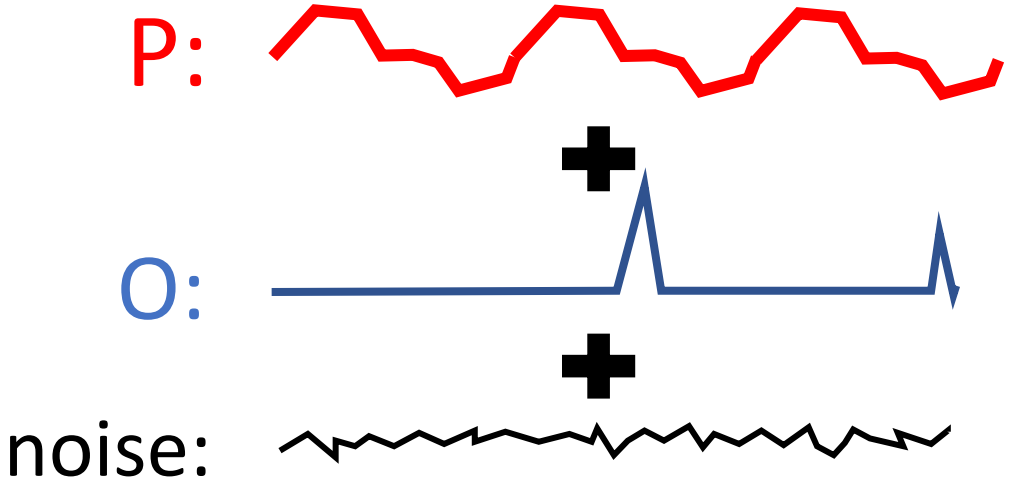
X

k



X: Periodic temporal behavior

How to impose structure on the temporal factors?



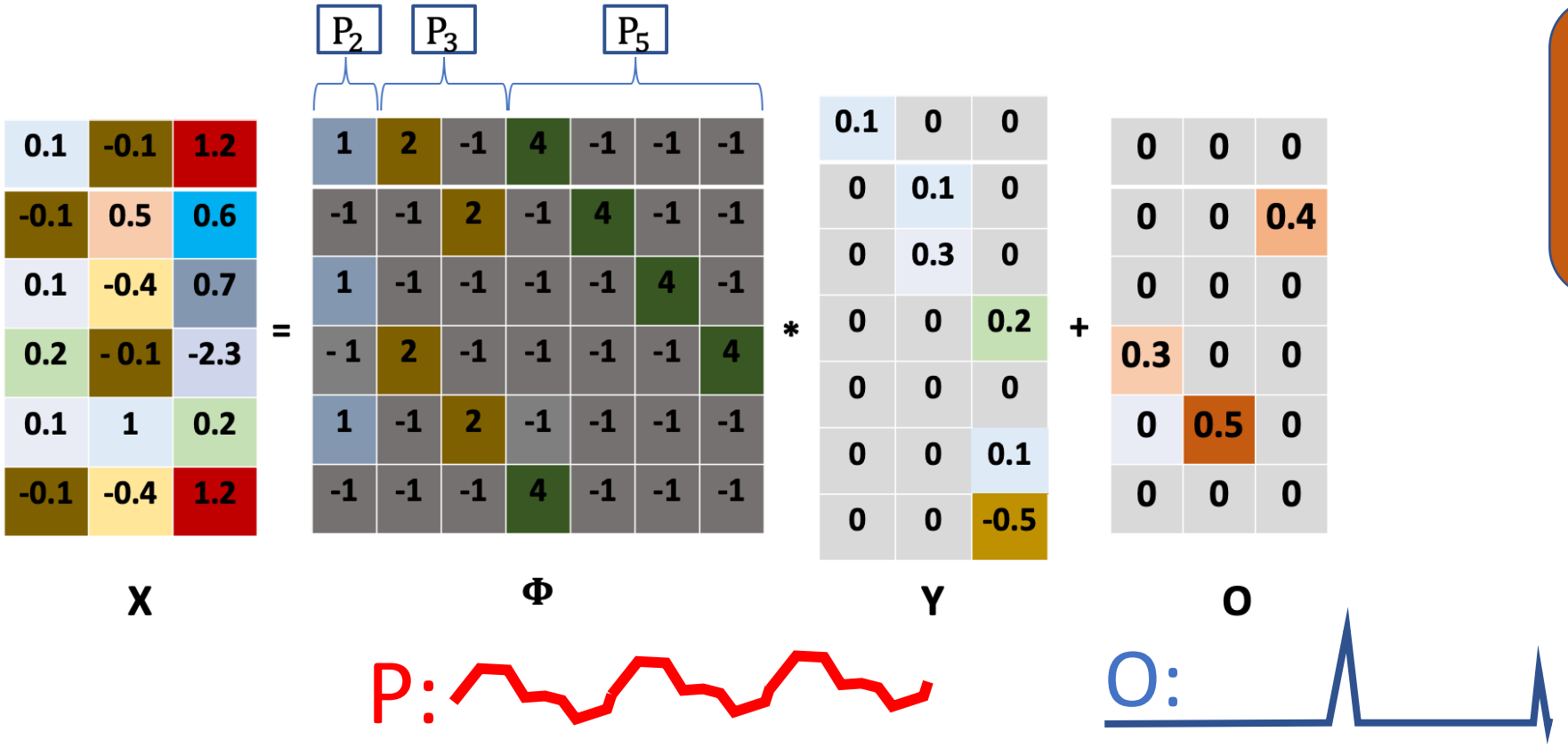
X = P + O + noise

Periodic component

Outlier component

Periodicity via a Ramanujan dictionary coding

Nested Periodic Matrix: Ramanujan Dictionary

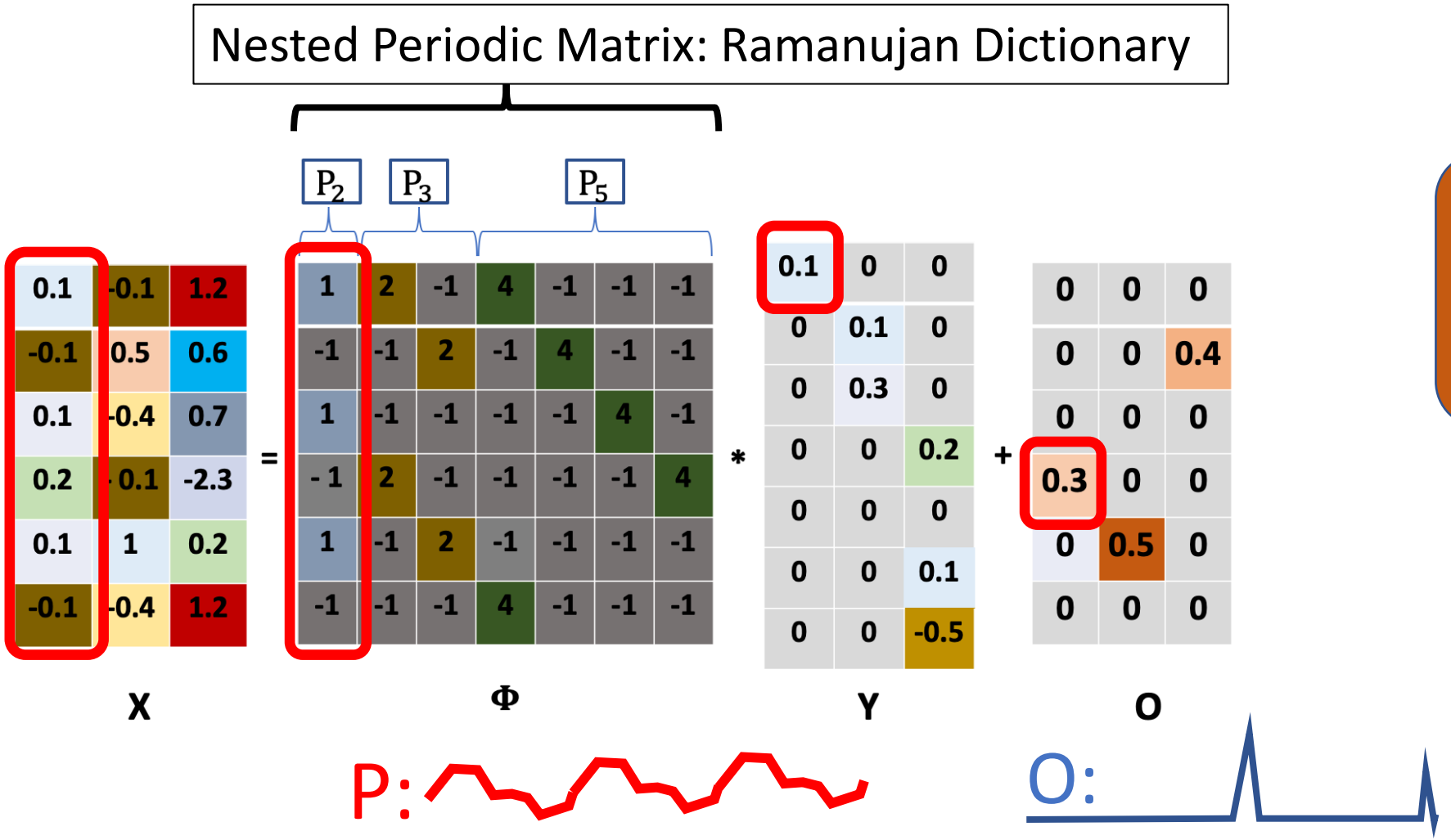


Details on NPMs and Ramanujan basis in the paper

P:

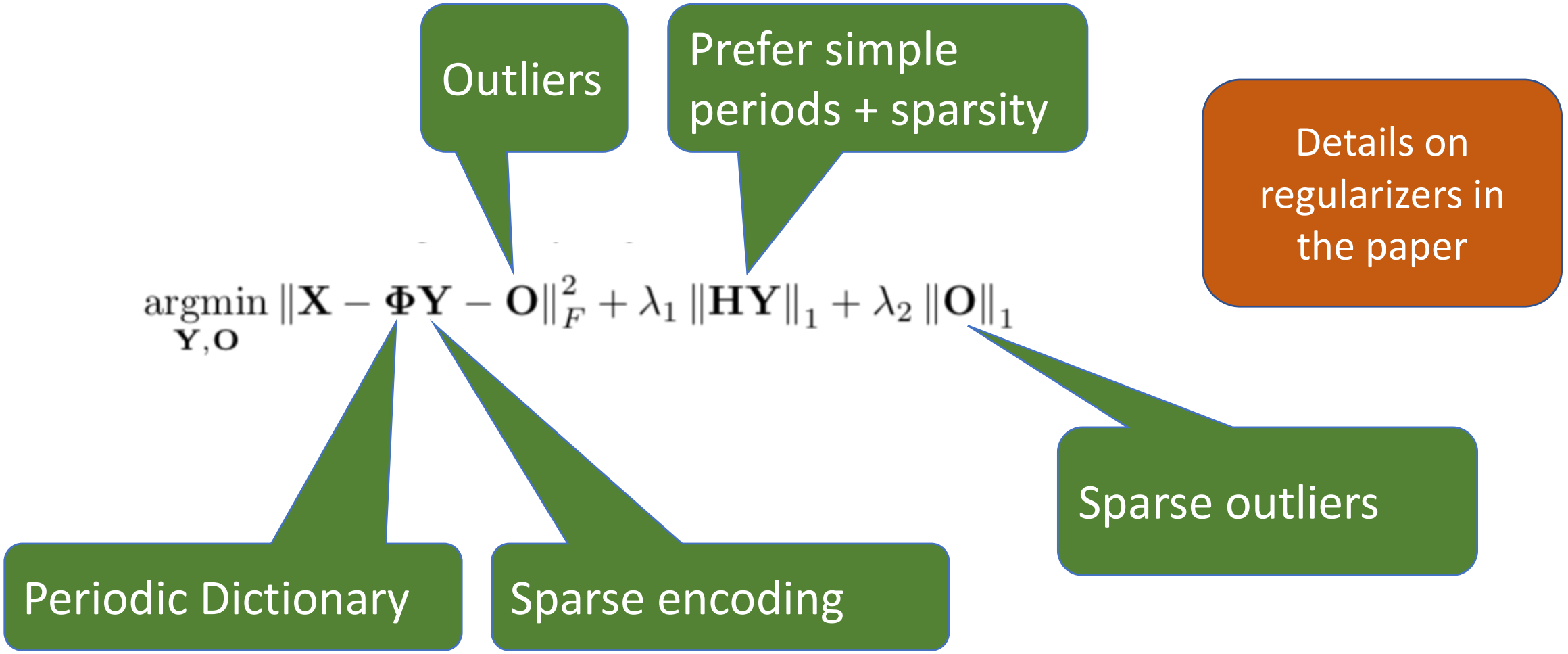
O:

Periodicity via a Ramanujan dictionary coding



Details on NPMs and Ramanujan basis in the paper

Temporal Regularization



Optimization for the PERCeIDs model

$$\underset{\mathbf{U},\mathbf{X},\mathbf{Y},\mathbf{O}}{\operatorname{argmin}} \underbrace{\frac{1}{2}\left\|\boldsymbol{\mathcal{W}} - [\mathbf{U},\mathbf{X}]\right\|_F^2}_{\text{Tensor Factorization}} + \underbrace{\lambda_0 \left\|\mathbf{X} - \boldsymbol{\Phi}\mathbf{Y} - \mathbf{O}\right\|_F^2 + \lambda_1 \left\|\mathbf{H}\mathbf{Y}\right\|_1 + \lambda_2 \left\|\mathbf{O}\right\|_1}_{\text{Temporal Structure}}$$

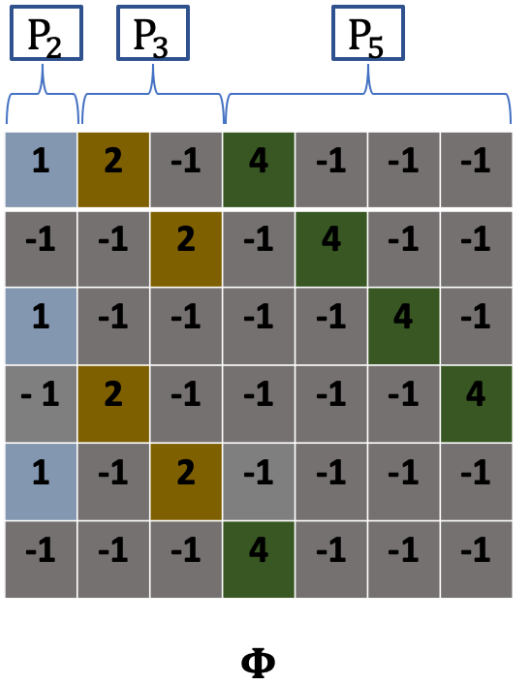
Highlights

- AOADMM solution (Non-differentiable components)
- High quality and reasonable speed in practice

How many candidate periods to consider? g_{max}

- Considering large periods increases the size of the dictionary (and computational cost)
- Automatic detection of the maximum period to explore for a given dataset

Details on g_{max} estimation in the paper



Experiments

Evaluation – data, metrics, baselines

	Statistics		
Dataset	$\lvert\mathcal{V}\rvert$	T	K
Synthetic	150	200	5
Football	115	1243	12
Reality Min.	94	8636	7
Reddit-Episode	242	8636	7
Reddit-TVshows	3538	1641	6

Ground truth communities available

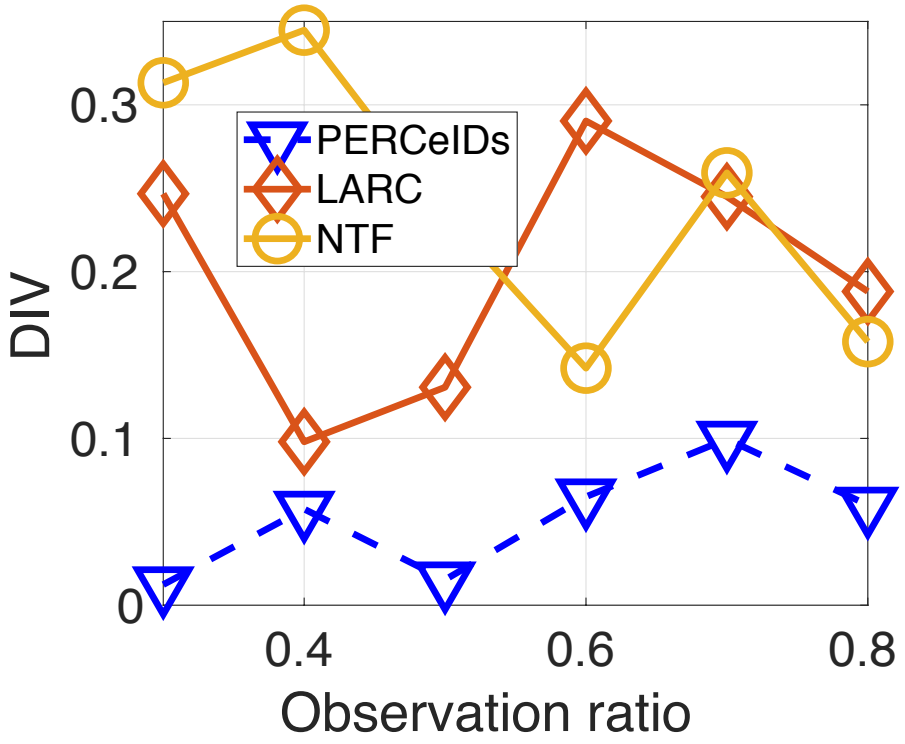
Metrics

- DIV - JS divergence of community weights (*lower better*)
- NMI – normalized mutual information (*higher better*)

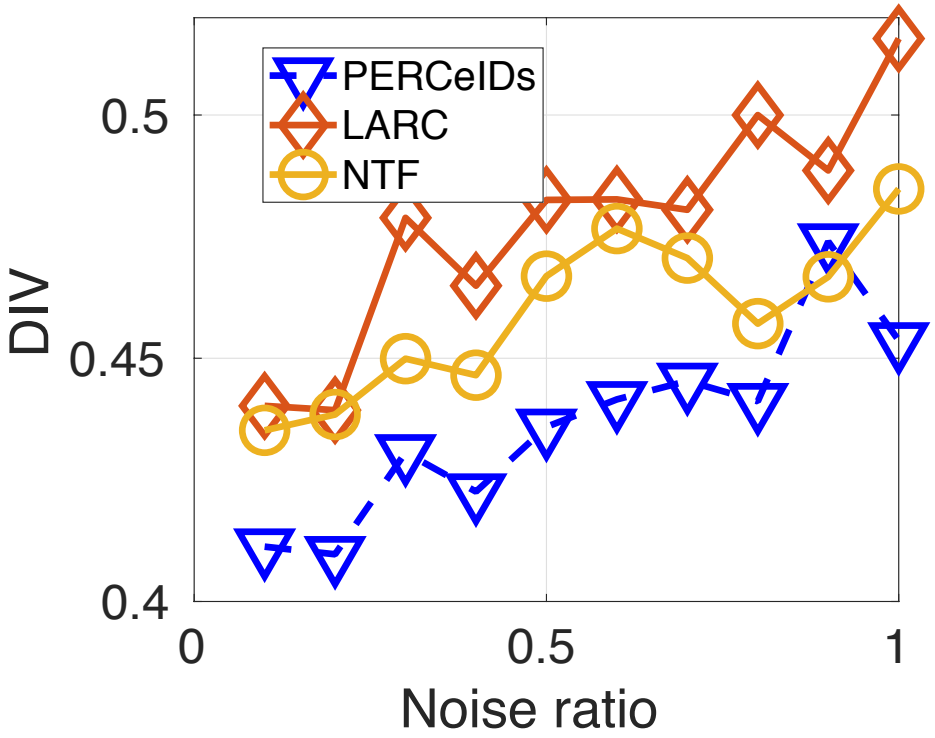
Baselines:

- LARC [KDD'18] (on/off temporal structure)
- TF [PlosOne'14] (no structure, vanilla TF)

Community quality (DIV lower is better)

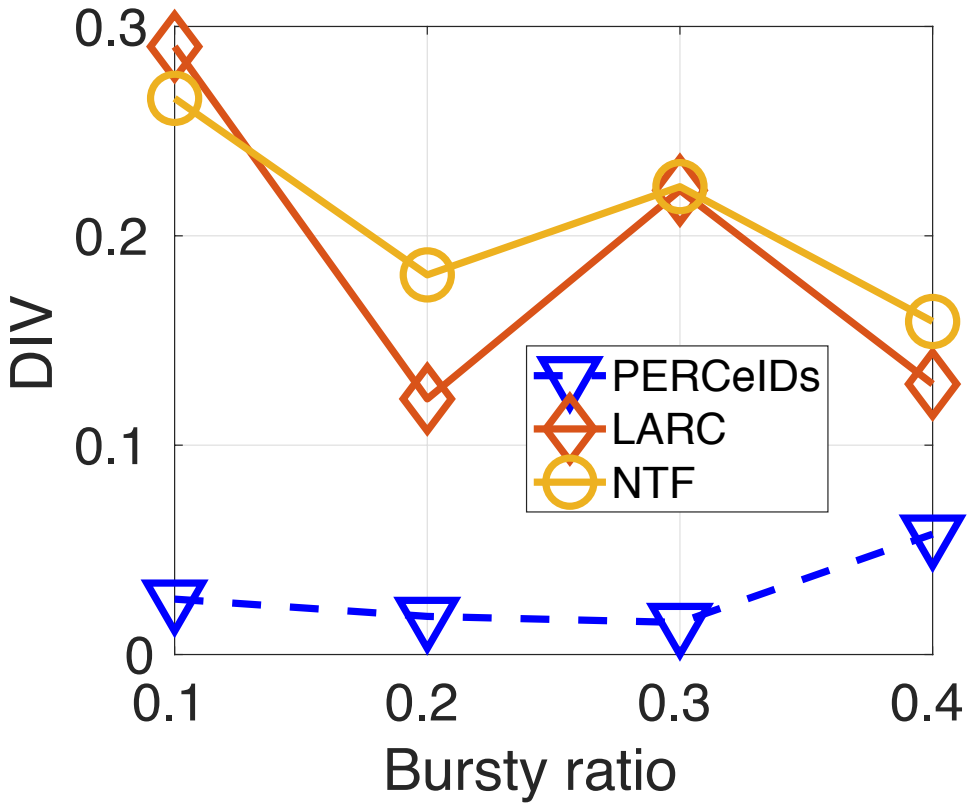


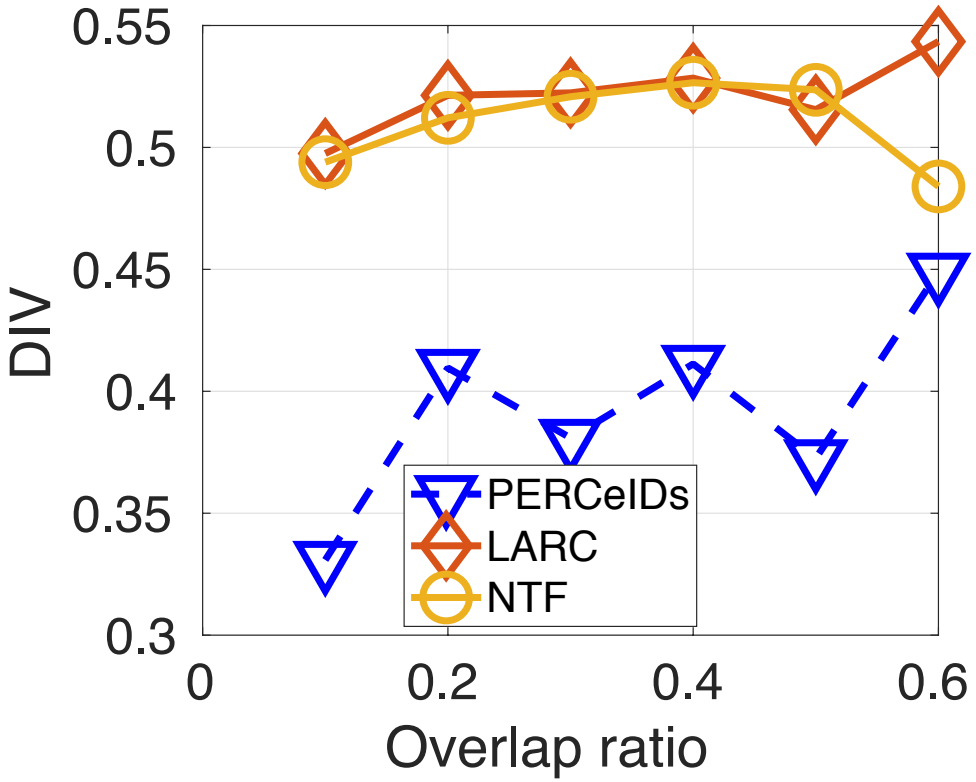
Increasing observation window



Increasing noise level

Community detection: outliers and overlap





Evaluation – real datasets summary

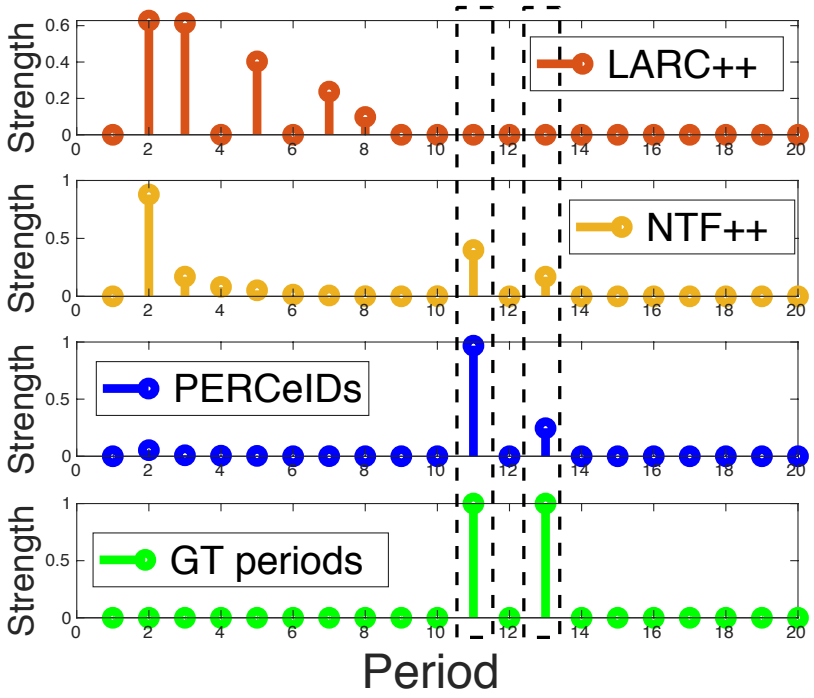
	Statistics			PERCeIDs				LARC [25]			NTF [23]		
Dataset	$\|\mathcal{V}\|$	T	K	$\hat{g}_{max}$	DIV	NMI	Time	DIV	NMI	Time	DIV	NMI	Time
Synthetic	150	200	5	17	**0.03**	**0.98**	5	0.30	0.87	3	0.28	0.82	3
Football	115	1243	12	26	**0**	**1**	5	0.008	0.91	3	0.14	0.77	3
Reality Min.	94	8636	7	23	**0.55**	**0.21**	20	0.65	0.17	40	0.80	0.06	7
Reddit-Episode	242	8636	7	31	**0.80**	**0.004**	15	0.88	0.003	30	0.94	0	10
Reddit-TVshows	3538	1641	6	39	**0.81**	0	76	0.82	0	70	0.96	0	48

TABLE II

DIV - JS divergence, a symmetric 0-1 divergence measure between community weights as distributions over nodes (lower better)

NMI – normalized mutual information, modified to allow for overlapping communities (higher better)

Evaluation - Period learning

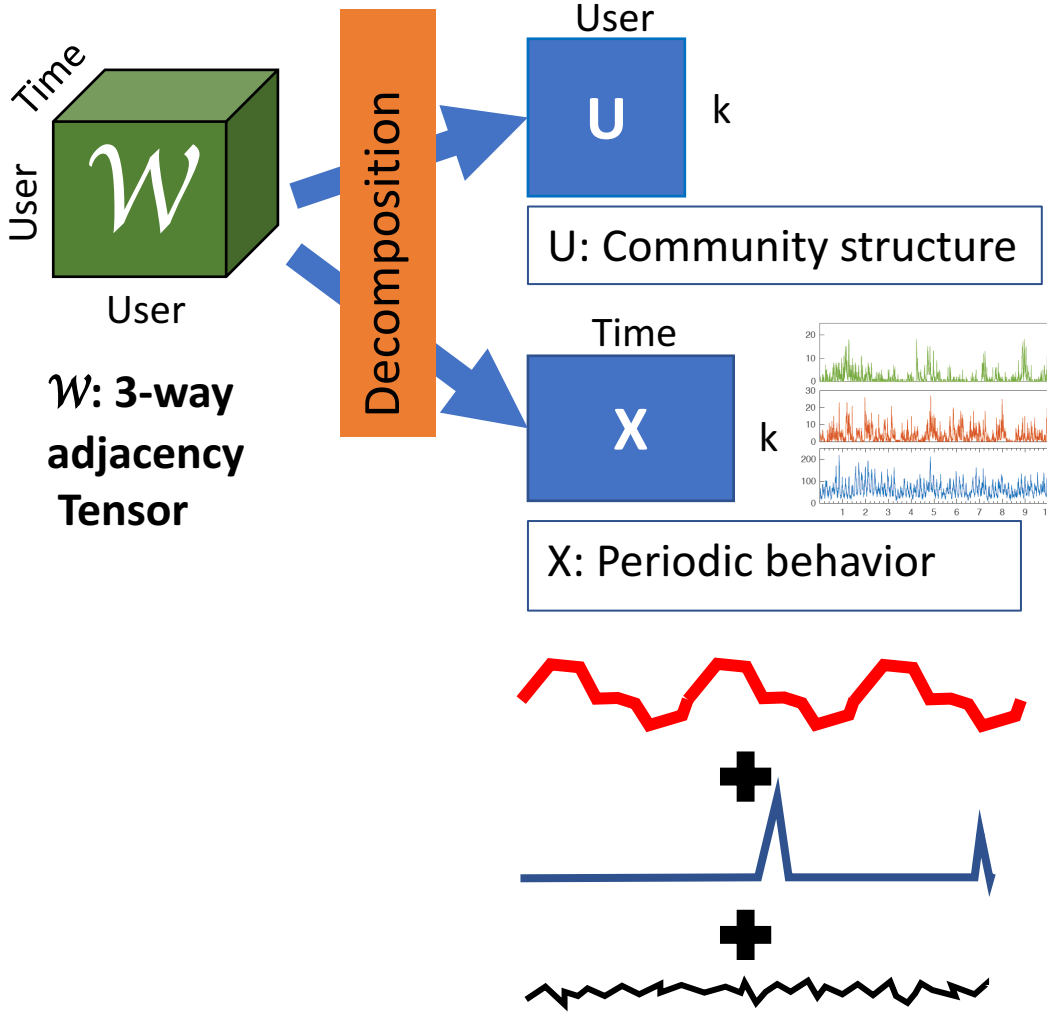


Temporal activities from LARC and NTF fail to capture the periodicity.

Additional results on outlier detection in the paper

Summary

- Novel problem formulation:
  - Tensor factorization with periodic temporal structure
- High quality on three tasks:
  - Community detection
  - Period estimation
  - Outlier time points detection
- Interpretability



UNIVERSITY AT ALBANY
State University of New York

DMNS LAB

Acknowledgments

Code:
https://github.com/LIN-ZHANG-ALPHA/PERCeIDs

On the job market

Lin Zhang

Alex Gorovits

Web Site:
https://lin-zhang-alpha.github.io
Email: lzhang22@albany.edu

Sponsors

NSF

Questions?

Extra

Optimizing the PERCeIDs model

Sub-problems for AOADMM

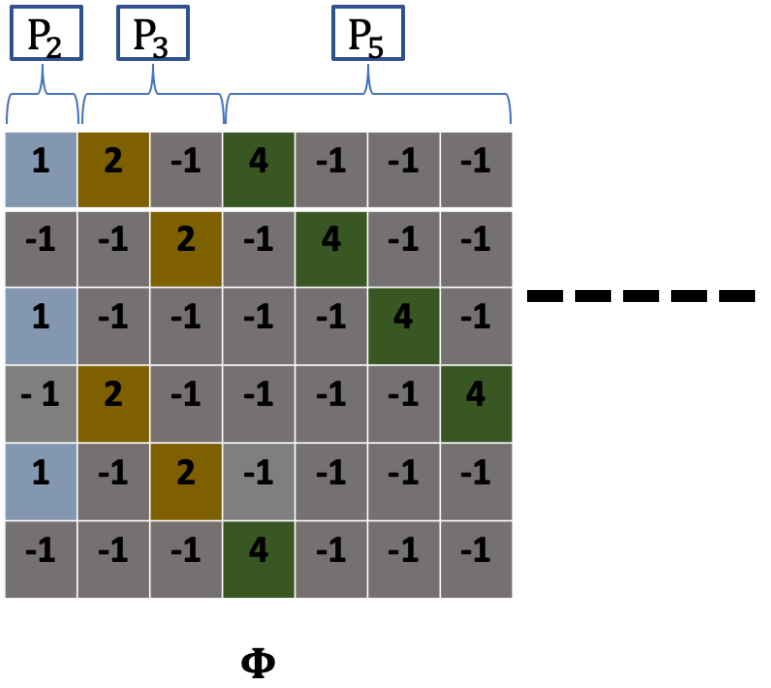
$$
\begin{cases}
\mathbf{U} : \underset{\mathbf{U}\geq 0}{\operatorname{argmin}} \ \frac{1}{2} \left\| \boldsymbol{\mathcal{W}} - [\mathbf{U}, \mathbf{X}] \right\|_F^2 & \text{(a)} \\
\mathbf{X} : \underset{\mathbf{X}\geq 0}{\operatorname{argmin}} \ \frac{1}{2} \left\| \boldsymbol{\mathcal{W}} - [\mathbf{U}, \mathbf{X}] \right\|_F^2 + \lambda_0 \left\| \mathbf{X} - \boldsymbol{\Phi}\mathbf{Y} - \mathbf{O} \right\|_F^2 & \text{(b)} \\
\mathbf{Y} : \underset{\mathbf{Y}}{\operatorname{argmin}} \ \lambda_0 \left\| \mathbf{X} - \boldsymbol{\Phi}\mathbf{Y} - \mathbf{O} \right\|_F^2 + \lambda_1 \left\| \mathbf{H}\mathbf{Y} \right\|_1 & \text{(c)} \\
\mathbf{O} : \underset{\mathbf{O}}{\operatorname{argmin}} \ \lambda_0 \left\| \mathbf{X} - \boldsymbol{\Phi}\mathbf{Y} - \mathbf{O} \right\|_F^2 + \lambda_2 \left\| \mathbf{O} \right\|_1 & \text{(d)}
\end{cases}
$$

How many candidate periods to consider? g_{max}

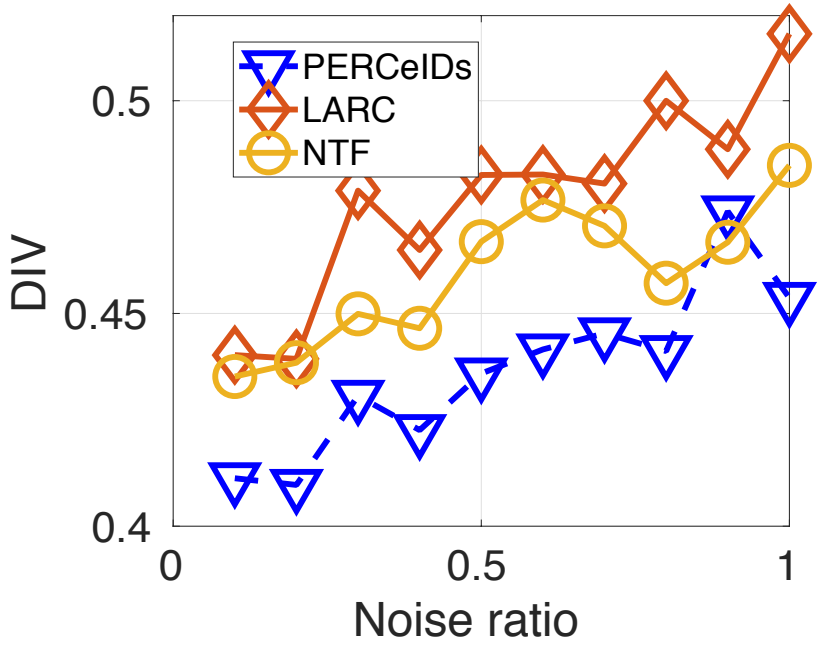
1. Reduce noise by using RPCA
2. Auto-correlation based estimation

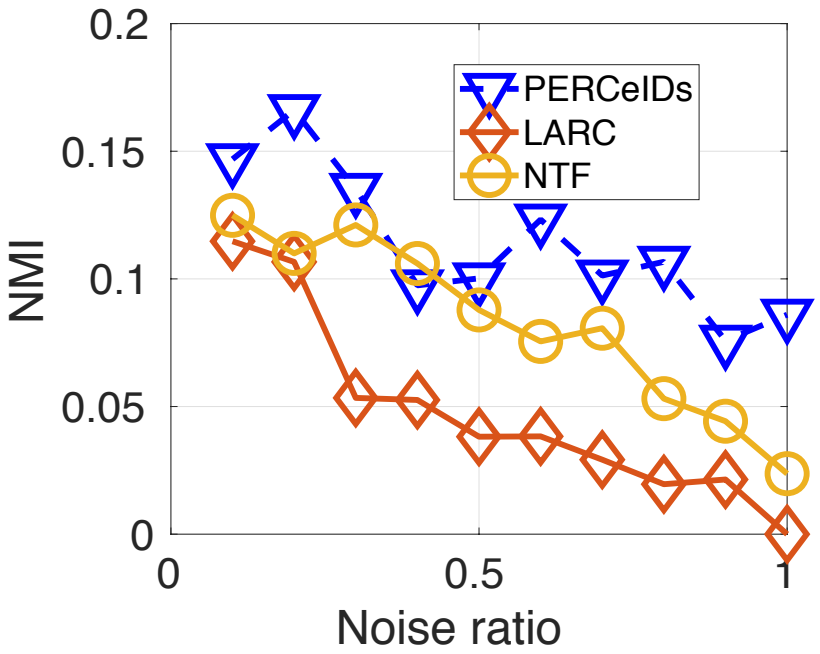
Algorithm 2: Estimate g_{max}

Input: Observations $\mathcal{W}$
Output: Maximum period g_{max}
Compute $\mathbf{X}$ by NTF on $\mathcal{W}$;
Compute covariance matrix $\mathbf{E} = \mathbf{X}\mathbf{X}^T$;
Separate signal from noise covariance: $[\mathbf{E}_s, \mathbf{E}_n]$ = RPCA($\mathbf{E}$);
$\mathbf{E}_s = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^T$;
$\mathbf{\Pi}_{opt} = \mathbf{M}\mathbf{\Lambda}\left(\mathbf{\Lambda} + \mu\mathbf{M}^T\mathbf{E}_n\mathbf{M}\right)^{-1}\mathbf{M}^T$;
$\mathbf{X}_s = \mathbf{\Pi}_{opt}\mathbf{X}$;
P_{range} = autocorr($\mathbf{X}_s$);
$g_{max} = max(P_{range})$.

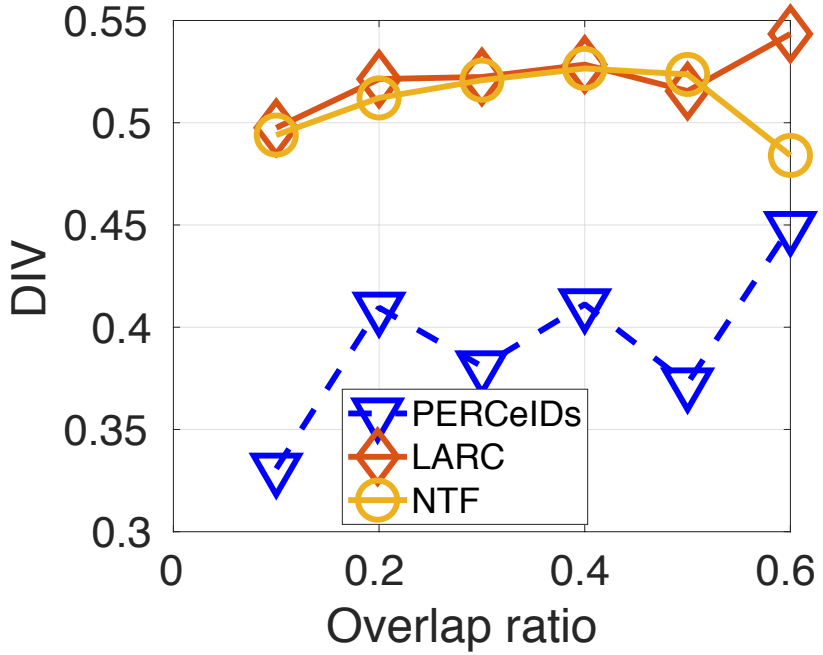


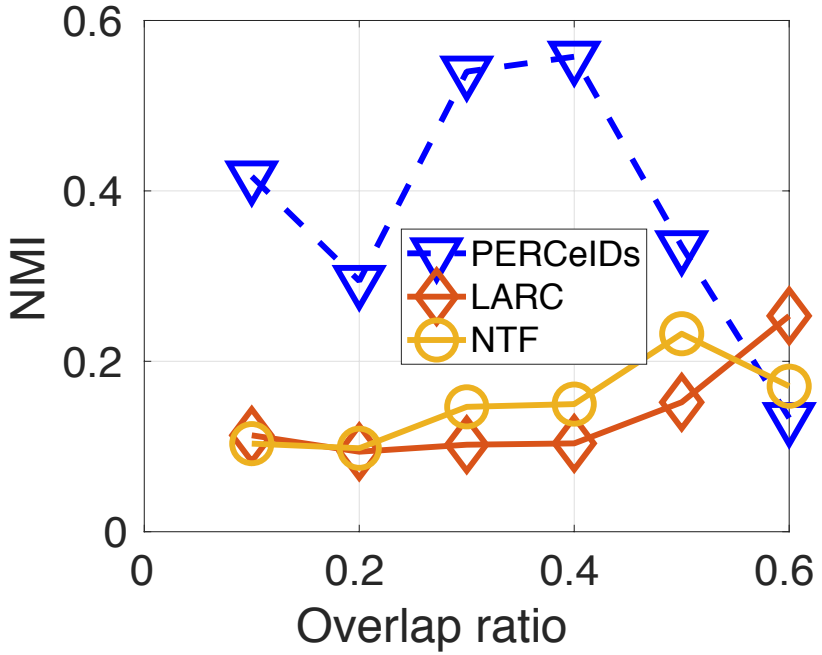
Evaluation – PERCeIDs retains quality through varying noise ratio



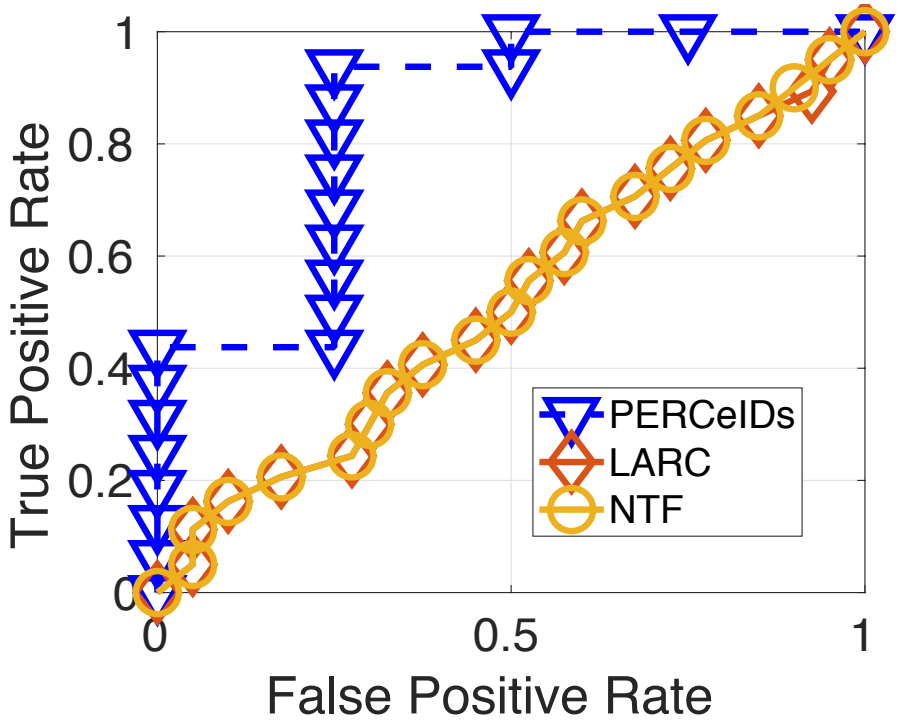


Evaluation – PERCeIDs retains quality through overlap ratio





Evaluation – outlier detection



Results of LARC and NTF are same as random guess because these two have no counterpart for outliers detection in their models.