ACM The Web Conference 2023
User Modeling and Personalization Track

Fine-tuning Partition-aware Item Similarities for Efficient and Scalable Recommendation

Tianjun Wei, Jianghong Ma, Tommy W.S. Chow

May 4, 2023
Austin, TX, USA

香港城市大學
City University of Hong Kong

Collaborative Filtering

Matrix Completion

- Completing the elements of the user-item interaction matrix that are not 1

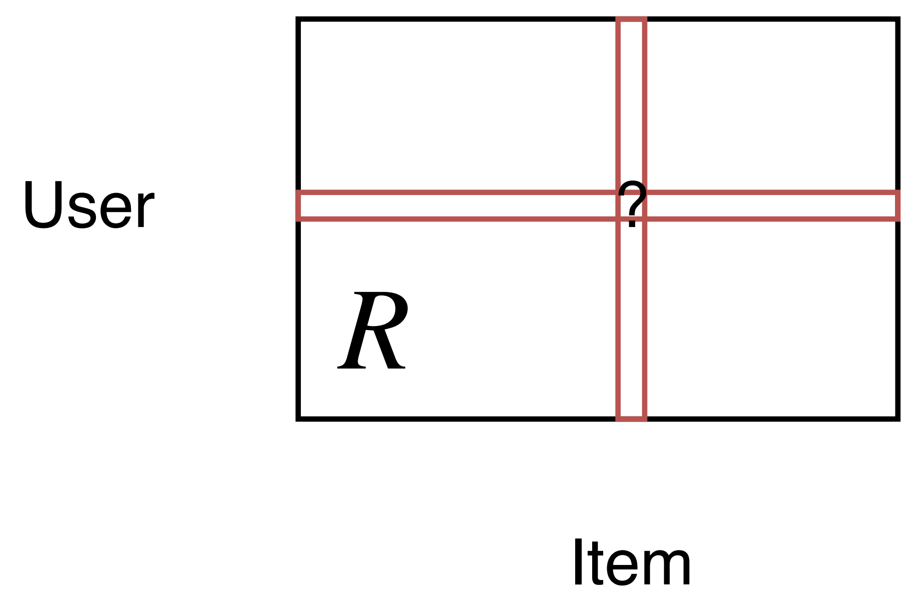
Link Prediction

- Predicting unconnected edges in user-item bipartite graphs

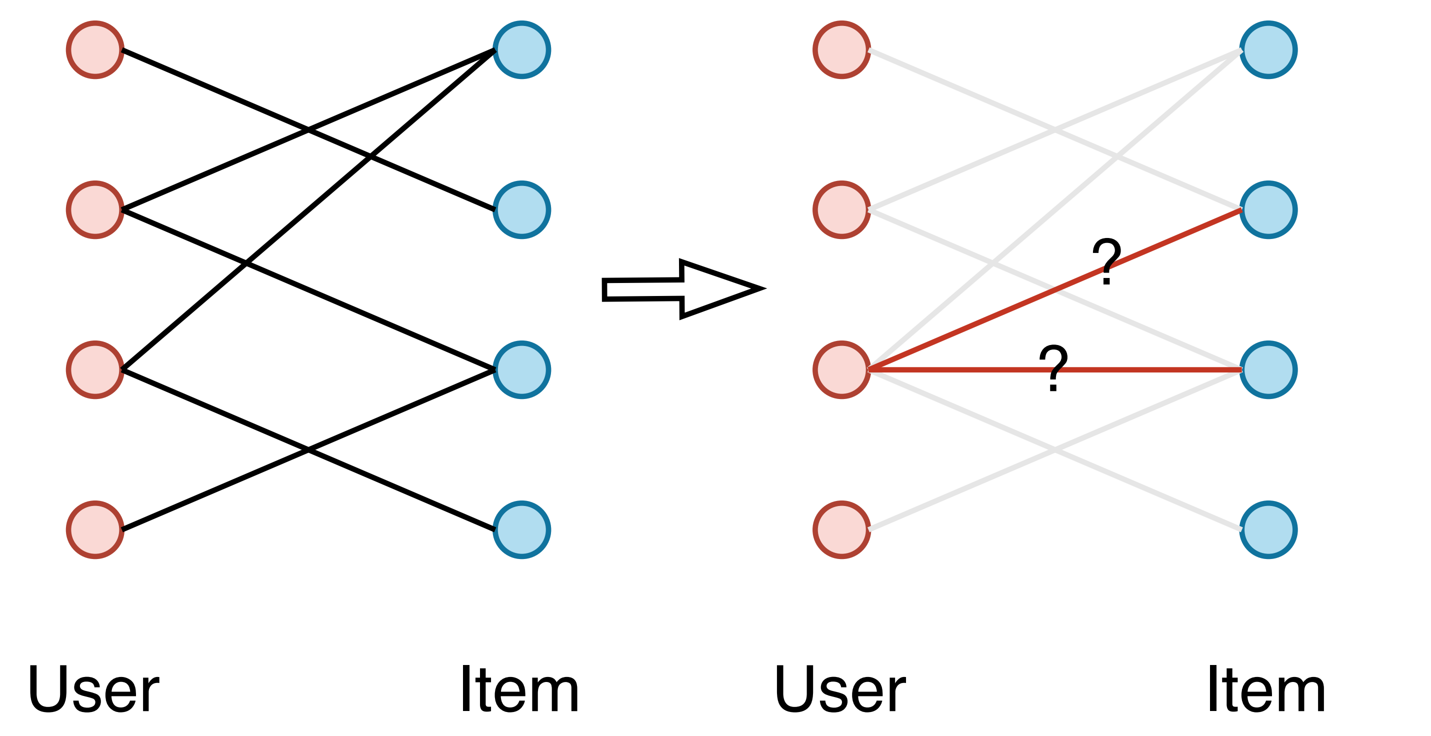
Nature of Recommendation Tasks

- Focus more on ranking of scores
- Focus more on top positions

Collaborative Filtering as a Matrix Completion Problem



Collaborative Filtering as a Link Prediction Problem



Type of Methods

Traditional Matrix Factorization

- Representing users and items with fixed-length vectors
- Both a trainable parameters ($N \propto |U| + |I|$)

Asymmetric Matrix Factorization

- Training only item vectors
- Representing users with the aggregation of item vectors
- No need to train when new users are added

Linear Autoencoder

- Full-rank Extension of asymmetric matrix factorization
- Adding sparse constraint to low the cost of storage and inference
- SLIM (ICDM 2011), EDLAE (Netflix, NIPS 2020)

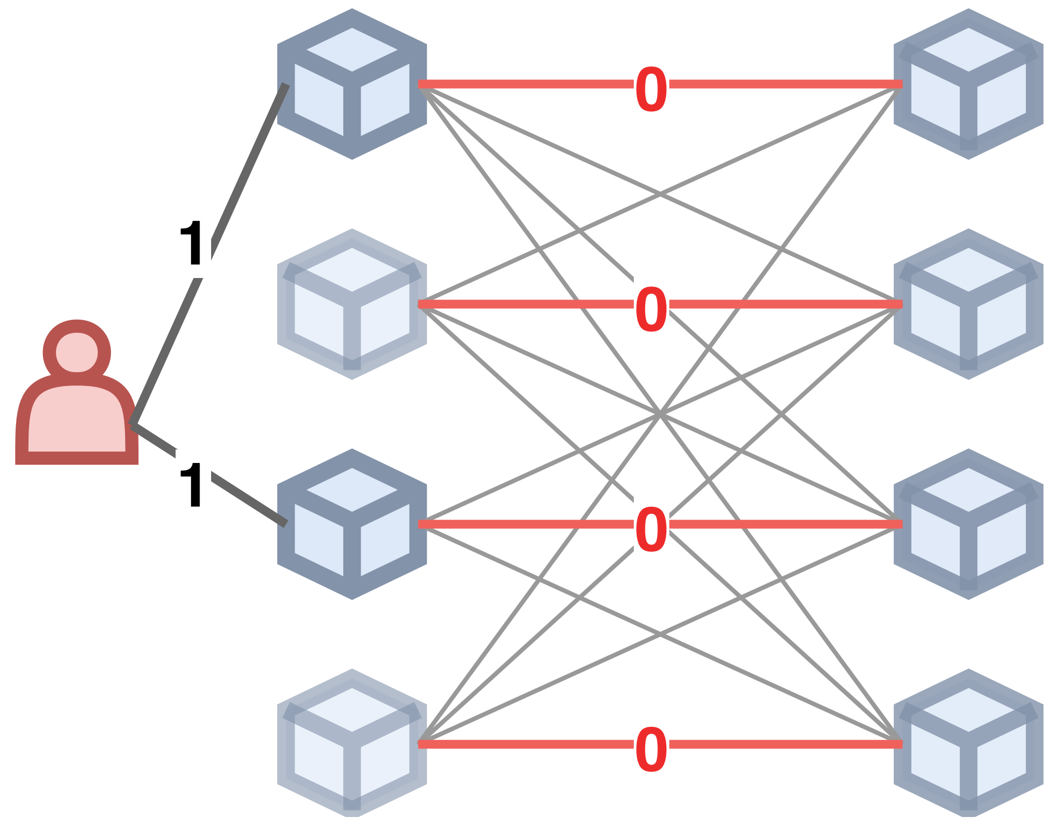
Ning, Xia, and George Karypis. "Slim: Sparse linear methods for top-n recommender systems." *ICDM 2011.*
Steck, Harald. "Autoencoders that don't overfit towards the identity." *NIPS 2020.*

Linear Autoencoder

Problem Formulation

$$\arg\min_{C} \|R - RC\|_F^2 + \frac{\theta_2}{2}\|C\|_F^2 + \theta_1\|C\|_1, \quad s.t.\, diag(C) = 0$$

- C — Weight matrix
- $\|R - RC\|_F^2$ — Linear autoencoder with hidden layer removed
- $\frac{\theta_2}{2}\|C\|_F^2$ — L2 Regularization
- $\theta_1\|C\|_1$ — L1 Regularization
- $diag(C) = 0$ — Diagonal-zero constraint (Avoid $C = I$)



Efficiency and Scalability Problem

- Computational complexity $O(|I|^2)$

Steck, Harald, et al. "Admm slim: Sparse recommendations for many users." *WWW 2020*.

Motivation - Start from GCN

LightGCN

- Graph convolutional layers with linear activation
- Increasing layers - low efficiency and over-smooth problem

$$E^{(l+1)} = (\tilde{D}^{-\frac{1}{2}} A \tilde{D}^{-\frac{1}{2}}) E^{(l)}$$

$$E = \frac{1}{L+1} \sum_{l=0}^{L} E^{(l)}$$

UltraGCN

- Approximation of infinite layers of LightGCN
- Convergence status ($E^{(l+1)} \approx E^{(l)}$)
- Convert graph convolution to the problem of maximizing the weighted inner product of vectors
  - User - Item interaction graph ($\tilde{R} = D_U^{-\frac{1}{2}} R D_I^{\frac{1}{2}}$)
  - Item - Item adjacency graph ($\tilde{Q} = \tilde{R}^T \tilde{R} / R^T R$)

$$e_u = \frac{1}{d_u + 1} e_u + \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{d_u + 1}\sqrt{d_i + 1}} e_i$$

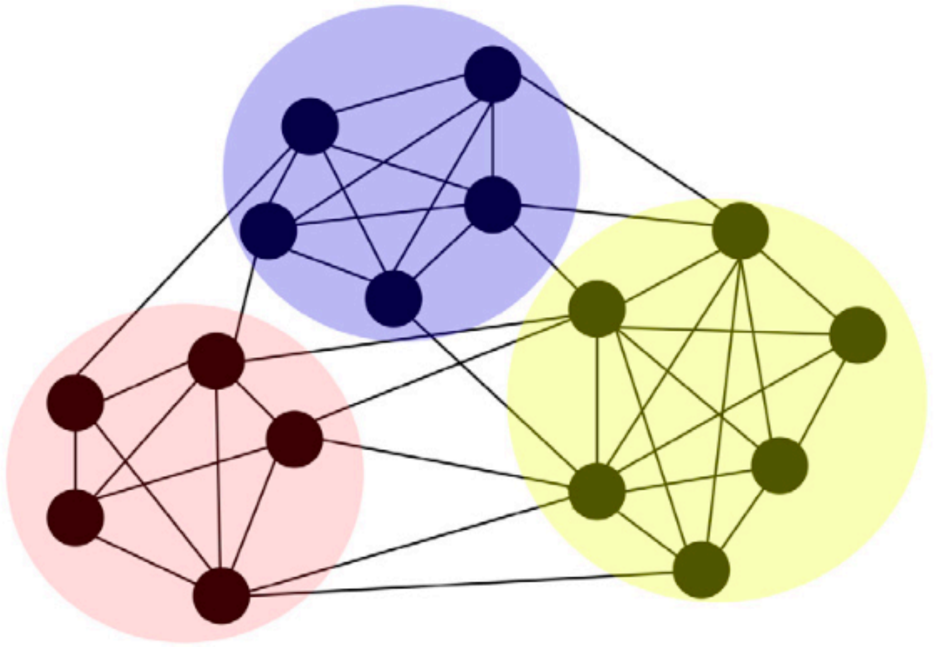
$$\beta_{u,i} = \frac{1}{d_u} \sqrt{\frac{d_u + 1}{d_i + 1}}$$

$$\omega_{i,j} = \frac{Q_{ij}}{\sum_k Q_{ik}} \cdot \frac{\sum_k Q_{jk}}{\sum_k Q_{ik} - Q_{ii}}$$

He, Xiangnan, et al. “LightGCN: Simplifying and powering graph convolution network for recommendation." *SIGIR 2020.*
Mao, Kelong, et al. "UltraGCN: ultra simplification of graph convolutional networks for recommendation." *CIKM 2021.*

Sampling in UltraGCN

- I - I adjacency graph is the 2-hop relations of U - I graph
- Much more denser than U - I graph



Graph Sampling

- Select Top-k neighbors in I - I graph

$$\arg\max_{j \in I}^{k} \ \omega_{i,j}$$

Graph Partitioning

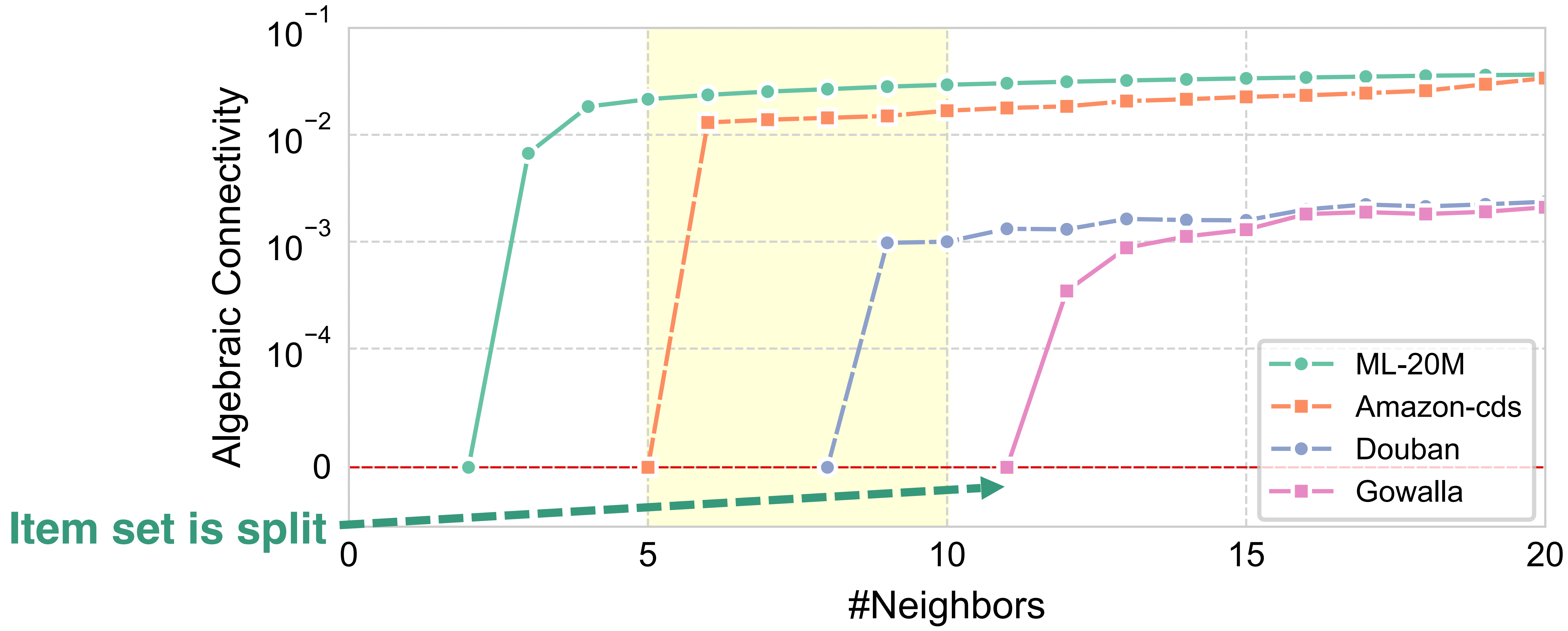
- Modularity maximization
- Graph sampling in UltraGCN can be considered as a local graph partitioning process

$$M(\mathcal{G}) = \frac{1}{2w} \sum_{i,j}^{\mathcal{E}} [A_{ij} - \frac{d_i d_j}{2w}] \mathbf{1}_{c_i = c_j}$$

$$M(\mathcal{G}_i) = \frac{1}{2d_i} \sum_{i,j} [\omega_{ij} - \frac{d_j}{2}] \mathbf{1}_{c_i = c_j} = \frac{1}{4d_i} \sum_{i,j} \omega_{ij} \mathbf{1}_{c_i = c_j}$$

Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E, 2004.*
Shen, Yifei, et al. "How powerful is graph convolution for recommendation?." *CIKM 2021*.

Community Structure in UltraGCN



Algebraic Connectivity (Fielder value) in the sampled I - I graph in UltraGCN

Our motivation

- Using global graph partitioning strategy to replace local one in UltraGCN
  - Neither real-world samples nor partitioning results are ideal
- Proposing new training strategy to alleviate information loss brought by partitioning

FPSR: Recursive Graph Partitioning

- Constructing full I - I graph requires high cost
- Using spectral partitioning without constructing the graph

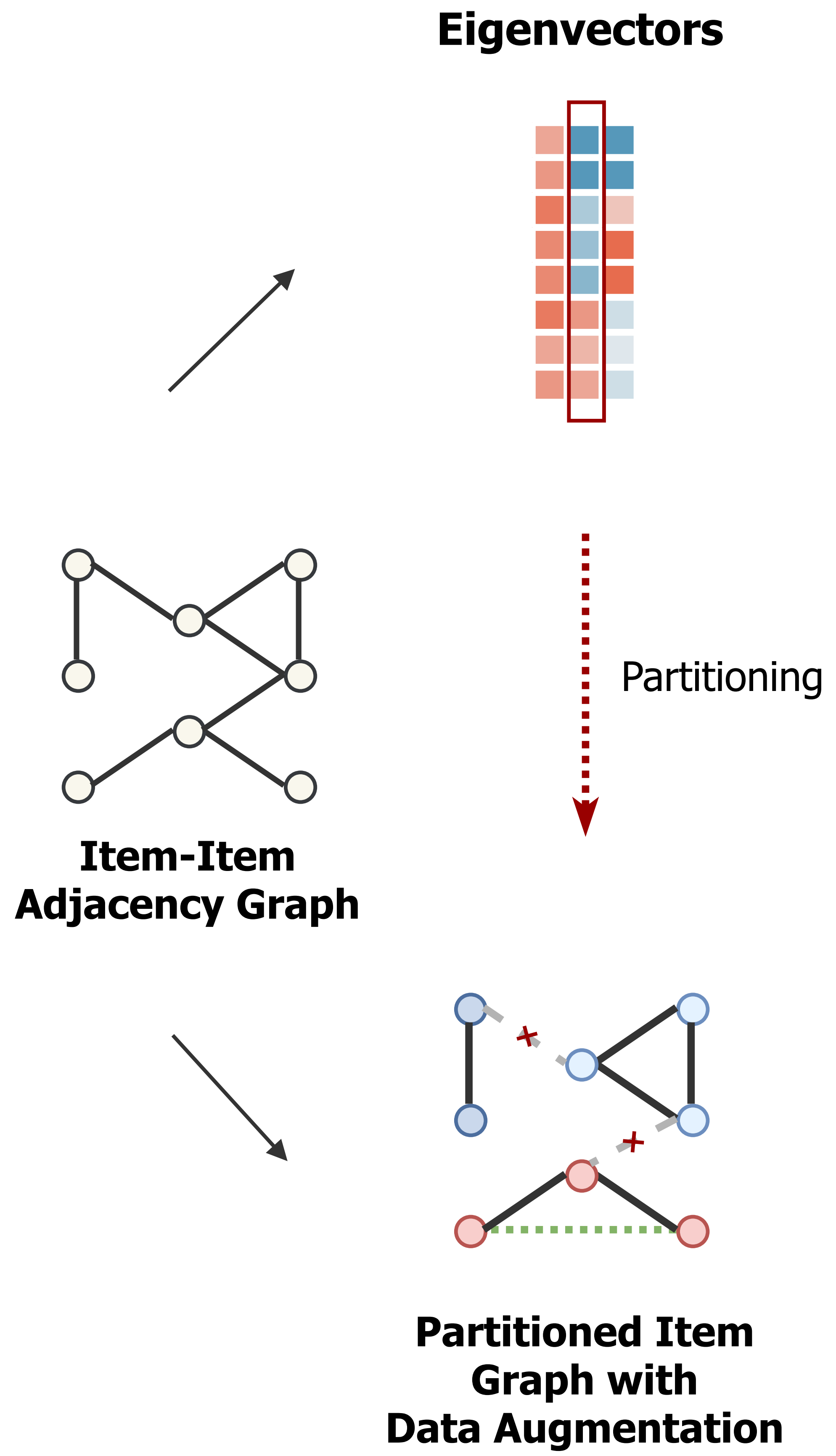
Fielder Vector

- Eigenvector corresponded to the second smallest eigenvectors in Laplacian matrix
- Equal to right singular vector corresponded to the second largest singular values in $\tilde{R}$
- Performing Truncated SVD on highly sparse $\tilde{R}$

$$i \in \begin{cases} I_1, & sign(\mathbf{v}_i) = +1, \\ I_2, & sign(\mathbf{v}_i) = -1. \end{cases}$$

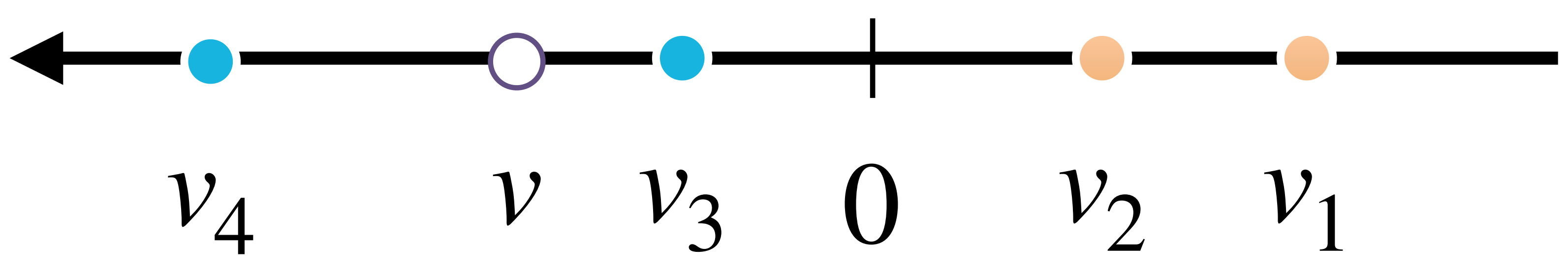
Recursive Partitioning

- Expecting the ratio between partition size and item size to be small
- Recursively performing partitioning process when partition size is not small enough



FPSR: Aggregate Global-level information

Fielder Vector



$$vv_4 > vv_3 > 0 > vv_2 > vv_1$$

- Implied item similarity information
- Adding to the learning of autoencoder through inner product
- Extending to Top-k singular vectors $V \in \mathbb{R}^{|I| \times k}$

Partition-aware Autoencoder

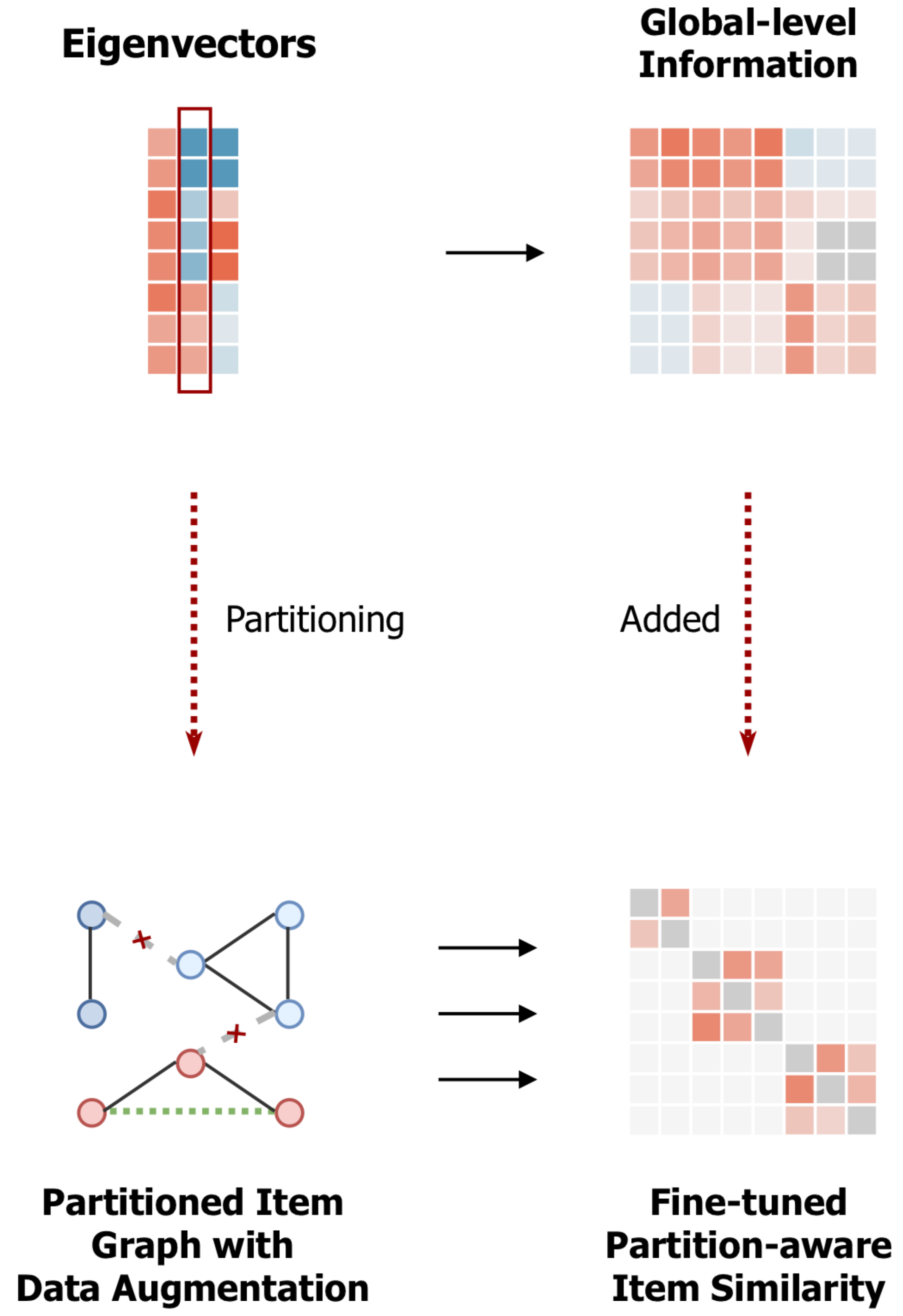
- V is the solution of the constraint optimization problem:

$$\arg\min_{U,V} \|\tilde{R} - UV^T\|_F^2, \; s.t. \; V^TV = I$$

- Aggregating global-level information

$$C = \underbrace{\lambda D_I^{-\frac{1}{2}} VV^T D_I^{\frac{1}{2}}}_{} + S$$

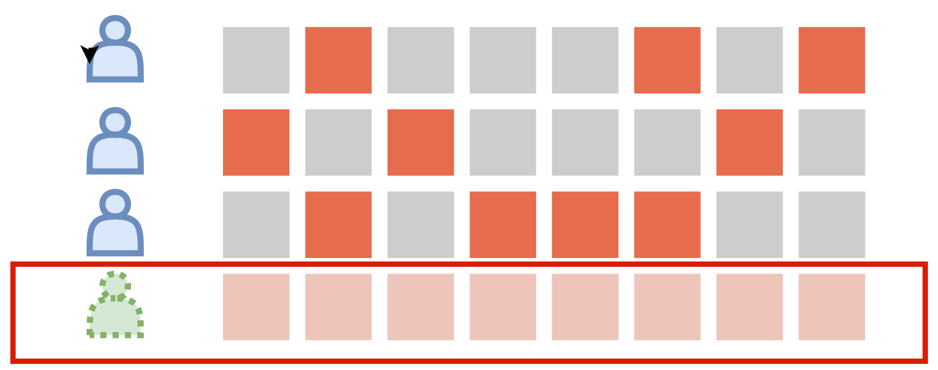
W



FPSR: Local Prior Knowledge of Partitions

- Items in the same partition can be considered having potential similarity
- Adding prior knowledge by inserting virtual users in R

$$\mathbf{1}_{n,i} = \begin{cases} 1, & \text{if } i \in I_n, \\ 0, & \text{otherwise.} \end{cases}$$

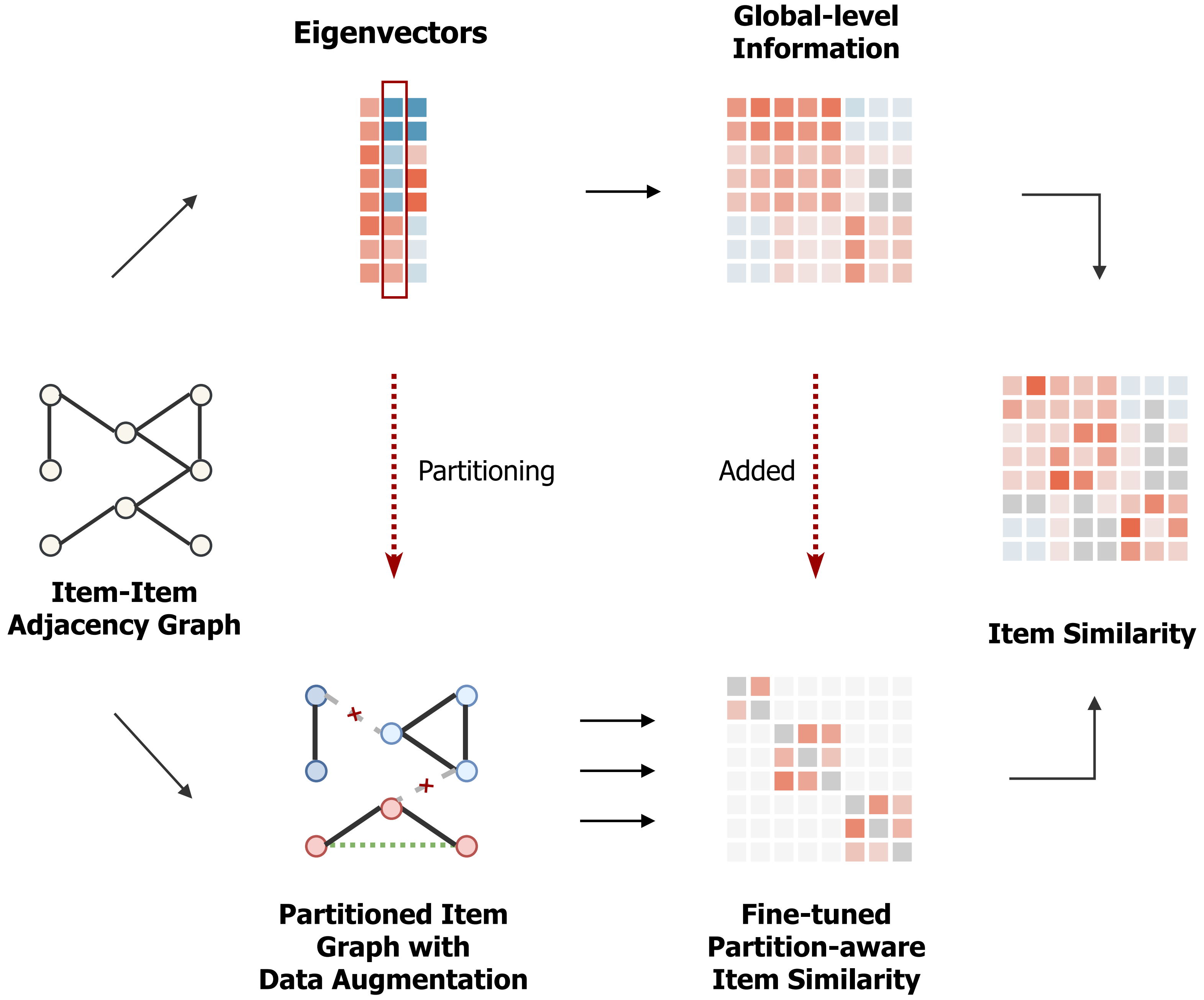


FPSR

$$\underset{S}{\operatorname{argmin}} \ \frac{1}{2}\|R - R(\lambda W + S)\|_F^2 + \frac{\theta_2}{2}\|D_I^{\frac{1}{2}}(\lambda W + S)\|_F^2$$

$$+\theta_1\|S\|_1 + \sum_n \frac{\eta}{2}\|\mathbf{1}_n^T - \mathbf{1}_n^T(\lambda W + S)\|_F^2$$

$$s.t. \ \ diag(S) = 0, \ S \geq 0, \ S_{ij|\mathcal{G}(i)\neq\mathcal{G}(j)} = 0,$$



Optimization

Alternating Direction Method of Multipliers (ADMM)

$$\underset{Z_n, S_n}{\operatorname{argmin}} \frac{1}{2} tr(Z_n^T \hat{Q} Z_n) - tr((I - \lambda W_n^T)\hat{Q} Z_n) + \theta_1 \|S_n\|_1$$
$$s.t.\ diag(Z_n) = 0,\ S_n \geq 0,\ Z_n = S_n,$$

Virtual users are not really inserted to R, but inserted by adding η

$$\hat{Q} = R_n^T R_n + \theta_2 D_I + \eta$$

Updating Rules

$$Z_n^{(t+1)} = (\hat{Q} + \rho I)^{-1}(\hat{Q}(I - \lambda W_n) + \rho(S_n^{(t)} - \Phi_n^{(t)}) - diagMat(\mu)),$$
$$S_n^{(t+1)} = (Z_n^{(t+1)} + \Phi_n^{(t)} - \frac{\theta_1}{\rho})_+,$$
$$\Phi_n^{(t+1)} = \Phi_n^{(t)} + Z_n^{(t+1)} - S_n^{(t+1)}.$$

Lagrange Multiplier

$$\mu = diag((\hat{Q} + \rho I)^{-1}(\hat{Q}(I - \lambda W_n) + \rho(S_n^{(k)} - \Phi_n^{(k)}))) \oslash diag((\hat{Q} + \rho I)^{-1})$$

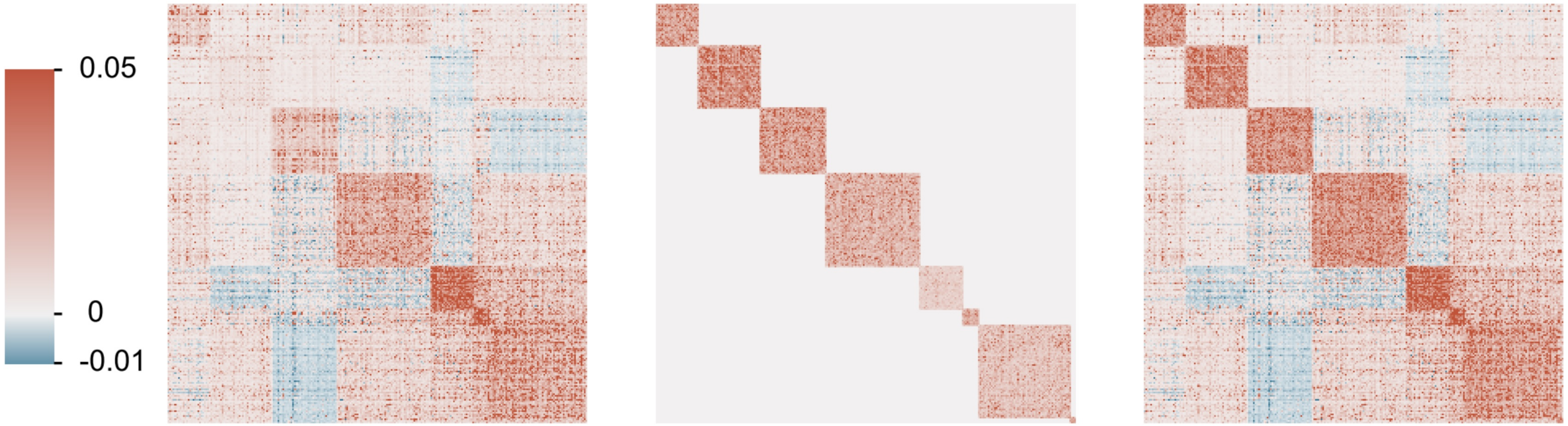
Experiments

- **Yelp2018** Dataset
  - 31,668 Users
  - 38,048 Items
  - 1,561,406 Interactions
- **Douban** Dataset
  - 13,024 Users
  - 22,347 Items
  - 792,062 Interactions

Dataset	Metric	GF-CF	LightGCN	SimGCL	UltraGCN	EASE	BISM	FPSR
Yelp2018	**Recall@20**	*0.0697*	*0.0653*	*0.0681*	*0.0683*	*0.0657*	*0.0662*	***0.0703***
	NDCG@20	*0.0571*	*0.0532*	*0.0556*	*0.0561*	*0.0552*	*0.0559*	***0.0584***
	#Params	-	*4.46M*	*4.46M*	*4.46M*	*1448M*	*191M*	*3.27M*
	Time (s)	*23*	*~74000*	*~2900*	*617*	*22*	*783*	*35*
Douban	**Recall@20**	*0.1719*	*0.1571*	*0.1699*	*0.1925*	*0.2038*	***0.2158***	*0.2095*
	NDCG@20	*0.1365*	*0.1206*	*0.1346*	*0.1556*	*0.1786*	*0.1889*	***0.1950***
	#Params	-	*5.04M*	*5.04M*	*5.04M*	*499M*	*158M*	*2.14M*
	Time (s)	*19*	*~22000*	*~2700*	*~5900*	*16*	*410*	*33*

Performance Comparison (Partial)

All experiments are conducted on with the same Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz machine with one Nvidia RTX A6000 GPU.



Global-level information | Fine-tuned item similarity | Weighted sum

Visualization of item similarity matrix learned in Yelp2018 dataset

Dataset	Amazon-cds		Douban		Gowalla		Yelp2018	
Metrics	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
w/o global-level information	0.1540	0.0873	0.2046	0.1909	0.1883	**0.1566**	0.0702	0.0582
w/o local prior knowledge	0.1542	0.0887	0.2085	0.1945	0.1785	0.1486	0.0662	0.0556
w/o regularization	0.1539	0.0870	0.2084	0.1943	0.1832	0.1501	0.0692	0.0574
FPSR	**0.1576**	**0.0896**	**0.2095**	**0.1950**	**0.1884**	**0.1566**	**0.0703**	**0.0584**

Ablation analysis

Detailed Analysis

Thanks!

GitHub: Joinn99/FPSR

Presenter: Tianjun Wei

City University of Hong Kong

tjwei2-c@my.cityu.edu.hk