

SDM'15 Vancouver, CAN

# Believe it Today or Tomorrow? Detecting Untrustworthy Information from Dynamic Multi-Source Data

Houping Xiao<sup>1</sup>, Yaliang Li<sup>1</sup>, Jing Gao<sup>1</sup>, Fei Wang<sup>2</sup>,  
Liang Ge<sup>3</sup>, Wei Fan<sup>4</sup>, Long Vu<sup>5</sup>, and Deepak Turaga<sup>5</sup>

<sup>1</sup> SUNY at Buffalo; <sup>2</sup> University of Connecticut;  
<sup>3</sup> Google; <sup>4</sup> Baidu Big Data lab; <sup>5</sup> IBM T.J. Watson

# Outline

- **Motivation**
- **Challenges**
- **Proposed Two-Step Framework**
  - Step-1: Joint Nonnegative Tensor Factorization
  - Step-2: Inconsistency Score calculation
- **Experiments**
- **Conclusions**

# Motivation

- **Multiple Information Sources**

- Example: Hotel ratings can be obtained from multiple websites, such as Priceline, Orbitz, and Tripadvisor

- **Question?**

- Which piece of information is trustworthy?
- Which object does receive reliable information?

- **Our solution**

- Calculate the degree of receiving inconsistent information across sources
- Lower degree of inconsistency – more reliable

# Motivating Example

# Motivating Example

The Jane Hotel, New York City

United States · New York (NY) · New York City · New York City Hotels

Search for a city, hotel, etc.

## The Jane

870 Reviews #298 of 463 Hotels in New York City

113 Jane Street, New York City, NY

Enter dates for best prices

Check In  Check Out

Show Prices

Compare best prices from top travel sites

Expedia Travelocity Booking.com and 6 more sites

Traveler photos Professional photos Browse nearby

★ ★ ★ ★ Budget West Village Manhattan

Overview Reviews (870) Photos (425) Location Amenities Q&A (5) Room Tips (159) Save

### 870 reviews from our community

Write a Review

Traveler rating	See reviews for	Rating summary
Excellent 285	Families 53	Location <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
Very good 351	Couples 158	Sleep Quality <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
Average 136	Solo 365	Rooms <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
Poor 58	Business 97	Service <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
Terrible 59		Value <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
		Cleanliness <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>

Related hotels...

- Hotel 31 Great Value!      1,143 Reviews New York City, NY Show Prices
- Broadway Hotel and Hostel      776 Reviews New York City, NY Show Prices
- Hotel 17 Great Value!      1,227 Reviews New York City, NY Show Prices

# Motivating Example

The image shows two overlapping screenshots of travel websites. The background is a TripAdvisor page for 'The Jane' hotel in New York City, featuring a 'Show Prices' button and a review summary. The foreground is a Priceline.com page for the same hotel, showing a price of \$109, a guest rating of 7.8/10, and a table of quality scores.

**tripadvisor** The Jane Hotel, New York

New York City > Hotels > Flights > Vacation Rentals

United States > New York (NY) > New York City > New York City

### The Jane

870 Reviews #298 of 403 Hotels

112 Jane Street, New York City, NY

Enter dates for best prices

Check In  Check Out

**Show Prices**

Compare best prices from top travel sites

Expedia Travelocity Booking.com and 8 more sites

Overview Reviews (870) Photos (425)

## 870 reviews from our community

Traveler rating	See reviews for
Excellent 285	Families
Very good 351	Couples
Average 136	Solo
Poor 58	Business
Terrible 59	

**priceline.com** Hotels Cars Flights Vacation Packages Sign In My Trips Help

## The Jane

Hotel Guides >> United States >> New York >> Hotels in New York City >> The Jane

**The Jane** **★★**  
Hotel Guest Rating: **7.8/10**  
112 Jane Street  
New York, NY 10014

see this hotel on a map | see more New York City Hotels

**Hotel Details** | **Guest Ratings & Reviews**

### Hotel Guest Rating for The Jane in New York City

See Most Popular hotels in New York City  
Based on 9 reviews from guests who stayed at the The Jane

Overall Quality Score		<b>7.8</b>
Hotel Cleanliness		<b>8.6</b>
Hotel Staff		<b>8.7</b>
Location of Hotel		<b>8.5</b>
Hotel Dining		<b>8.2</b>

from **\$109**

check-in

check-out

rooms

**check rates**

See this Hotel on a Map

#### Other Hotels Near The Jane

- Gansevoort Meatpacking **★★★★** from \$285 (0.2mi)
- The Maritime Hotel **★★★** from \$295 (0.4mi)
- Four Points By Sheraton Manhattan Soho Village **★★★★** from \$255 (0.8mi)
- Affinia Manhattan **★★★★** from \$141 (1.1mi)
- The Inn At Irving Place **★★★** (1.1mi)

# Motivating Example

The image displays three overlapping screenshots of travel websites: TripAdvisor, Priceline.com, and Yelp. A large yellow diagonal banner across the center reads "The Jane Hotel got inconsistent ratings from different websites!!!".

**Yelp Screenshot:** Shows "The Jane" with a 3.5 star rating and 43 reviews. A grey box highlights a discrepancy: "3.5 stars" is shown above a bar chart with 130 reviews, and "130" is shown next to the 1 star bar. Below the chart, text reads: "We calculate the overall star rating using only reviews that our automated software currently recommends. Learn more."

**Other Screenshots:** TripAdvisor shows 870 reviews and a "Show Prices" button. Priceline.com shows "The Jane" with a 4.5 star rating and 58 reviews. The bottom right shows a list of nearby hotels with their ratings and prices.

Website	Overall Rating	Number of Reviews
Yelp	3.5 stars	43
Yelp (Automated)	3.5 stars	130
Priceline.com	4.5 stars	58
TripAdvisor	4.5 stars	870

Category	Rating
Hotel Cleanliness	8.5
Hotel Staff	8.2
Location of Hotel	8.5
Hotel Dining	8.2

Hotel Name	Rating	Price
Affinia Manhattan	4 stars	from \$141
The Inn At Irving Place	3 stars	(1.1m)

# How to Find Inconsistent Ratings

- **Easy comparisons**

- Aggregate the ratings of all the users into average ratings will lose information
- Users are not matched across different platforms, so we are unable to compare ratings of each user

- **Our solution: Identify user groups and compare at the group level**

- In each source, users can be partitioned into groups so that users in the same group share similar rating patterns over objects
- The underlying user groups and the ratings given by each group should be consistent across sources



# Importance of Time

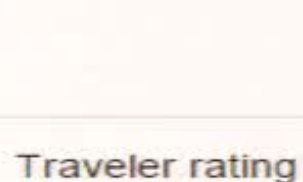
- **Observations**

- Multi-source data can continuously arrive with constantly changing distributions

- **Motivating Example**



## Santa Fe Station Hotel



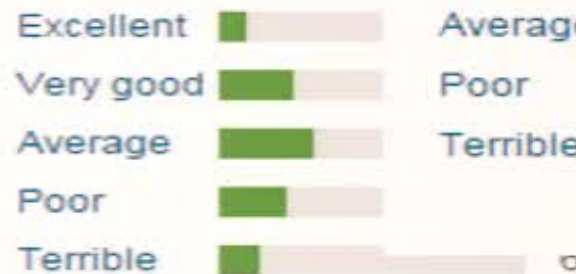
Mar-May

Traveler rating



Jun-Aug

Traveler rating



Sep-Nov

Traveler rating



Dec-Feb

**Ratings evolve over time!!!**

# Solutions

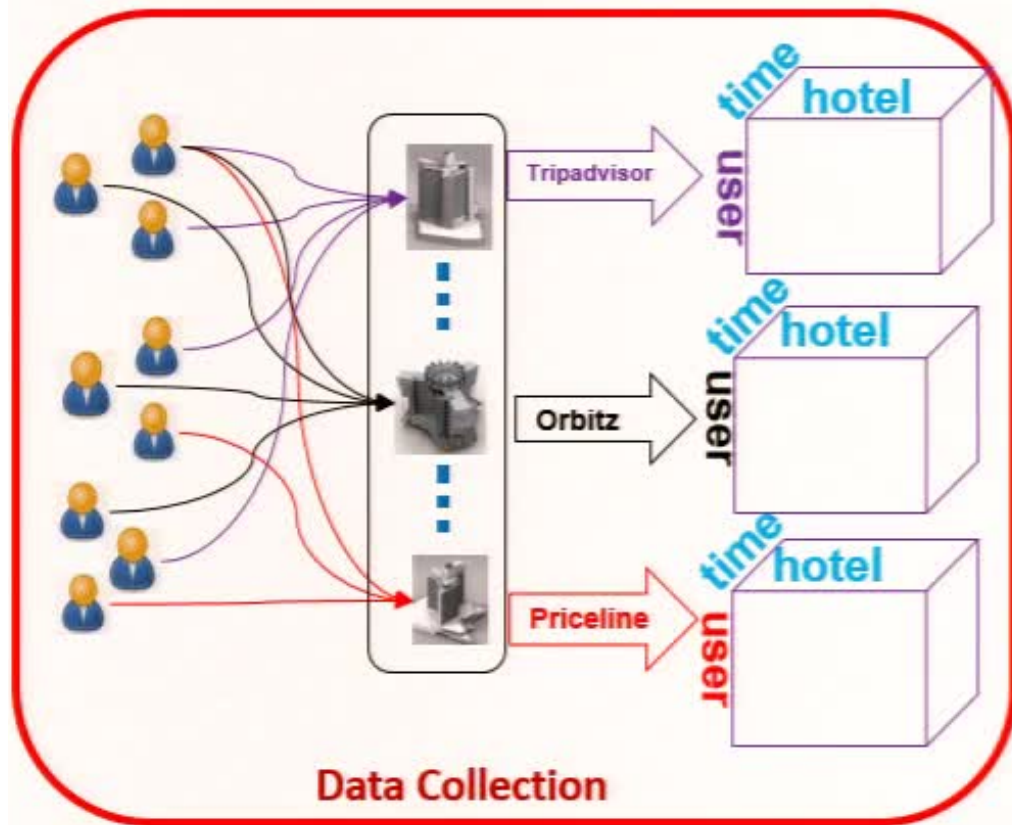
- **Baselines**

- Conduct separate modeling on each snapshot (Simple, but the temporal connection between timestamps is missing)

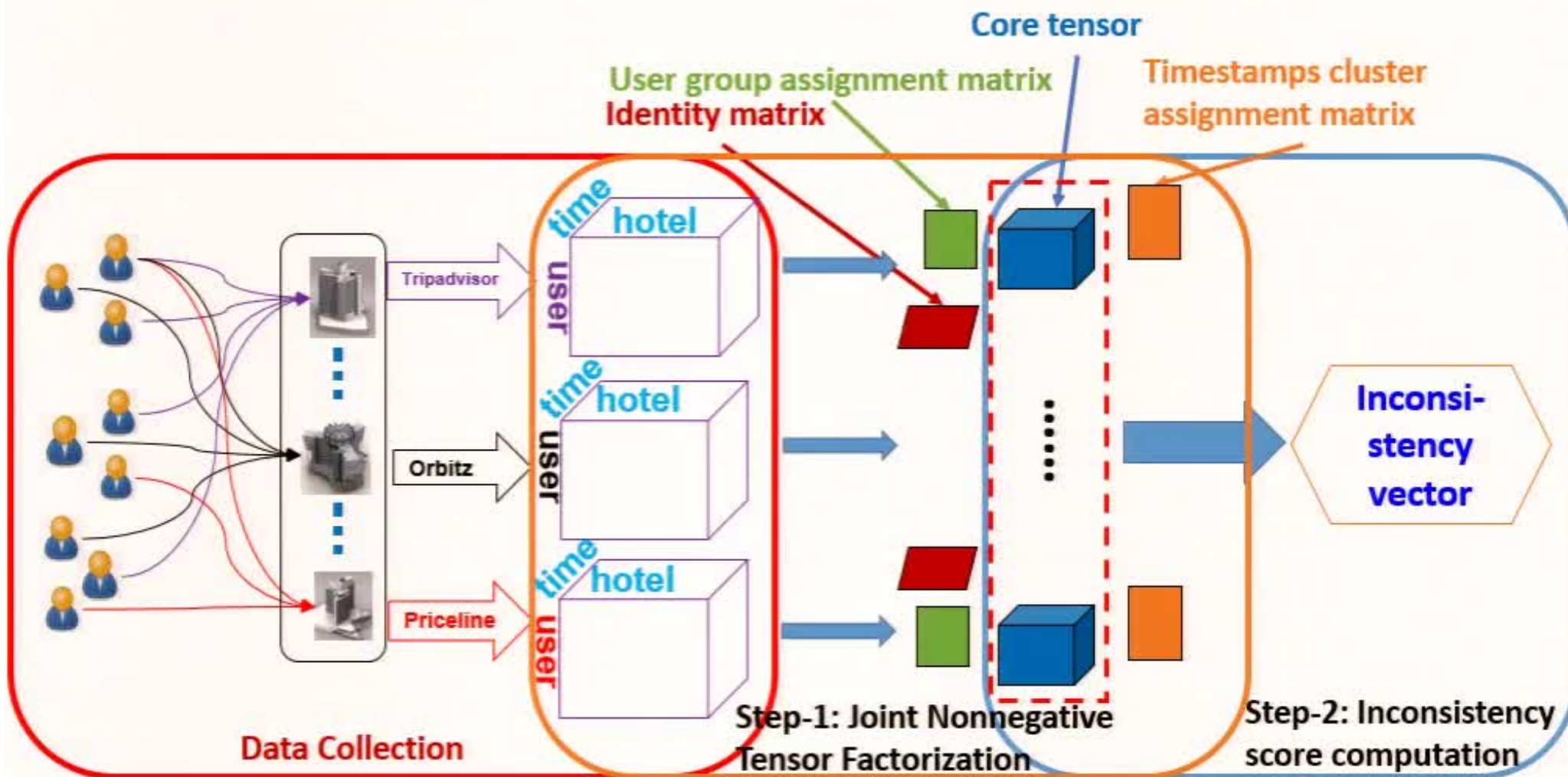
- **Our Solution**

- Consider the behavior at timestamp-cluster level (e.g. hotel ratings could change seasonally)
- In each source, timestamps can be clustered. Users' behavior at the same timestamp cluster should be similar

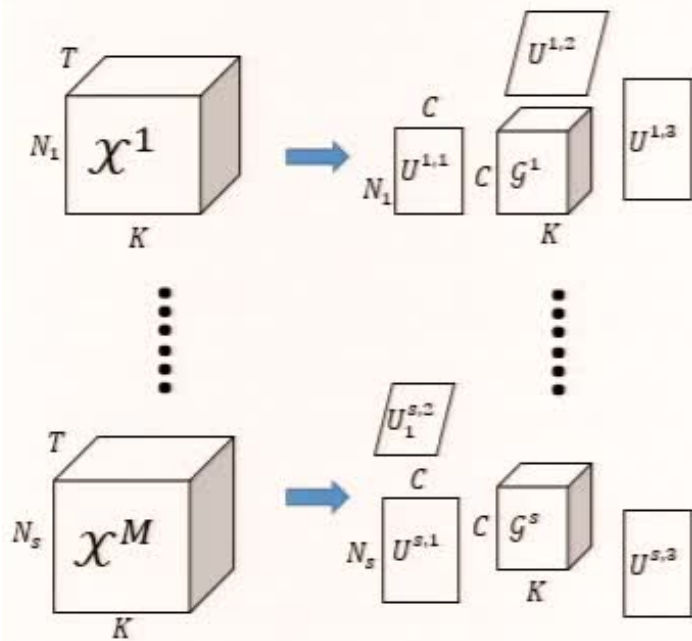
# Proposed Framework



# Proposed Framework

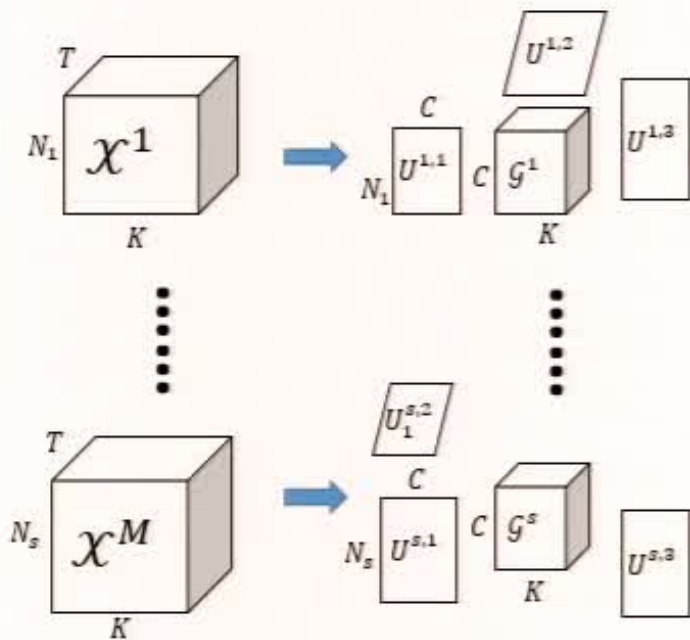


# Joint Nonnegative Tensor Factorization



$$\min_{\{\mathcal{G}^s, \{U^{s,i}\}\}_{s=1}^M \geq 0} \sum_{s=1}^M (\mathcal{L}(\tilde{\mathcal{X}}^s) + \alpha \Omega(\mathcal{G}^s, \mathcal{G}^*))$$

# Joint Nonnegative Tensor Factorization



$$\min_{\{\mathcal{G}^s, \{U^{s,i}\}\}_{s=1}^M \geq 0} \sum_{s=1}^M (\mathcal{L}(\tilde{\mathcal{X}}^s) + \alpha \Omega(\mathcal{G}^s, \mathcal{G}^*))$$

- $\mathcal{L}(\tilde{\mathcal{X}}^s) = \|\mathcal{X}^s - \mathcal{G}^s \times_i \{U^{s,i}\}\|_F^2$ , measures the factorization error of each tensor
- $\Omega(\mathcal{G}^s, \mathcal{G}^*) = \|\mathcal{G}^s - \mathcal{G}^*\|_F^2$ , where  $\mathcal{G}^* = \frac{1}{M} \sum_{s=1}^M \mathcal{G}^s$ , is a regularization term proposed to learn the consensus information
- $\alpha$ , is regularization parameter

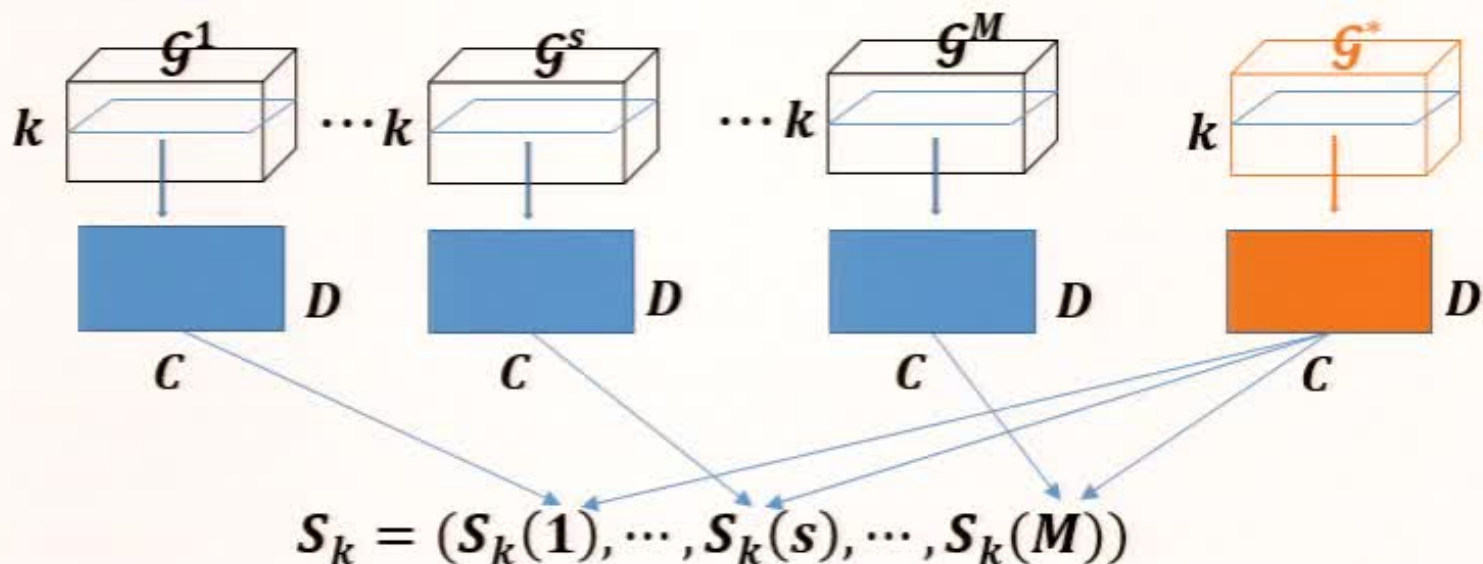
# Inconsistency Score Computation

- Inconsistency Score**

$$I_k = \|S_k - S_{median}\|^2,$$

where

- $S_k(s) = \text{similarity}(\mathcal{G}^s(:, k, :), \mathcal{G}^*((:, k, :)))$
- $S_{median} = \text{median}\{S_k, k = 1, \dots, K\}$



# Streaming Data

- **Observation:**

- Multi-source data continuously arrives

- **Solution:**

- Step 1: obtain  $\{U_o^{(s,T,i)}\}$  at time  $T$  based on  $\{\mathcal{G}^{s,T-1}\}$

$$\min_{\{U^{(s,i,T)}\}} \sum_{s=1}^M \left\| \mathcal{X}^{s,T} - \mathcal{G}^{s,T-1} \times_i \{U^{(s,i,T)}\} \right\|_F^2$$

- Step 2: use  $\{U_o^{(s,T,i)}\}$  to obtain  $\{\mathcal{G}^{s,T}\}$  at time  $T$

$$\min_{\{\mathcal{G}^{s,T}\}} \sum_{s=1}^M \sum_{t=1}^T \left\| \mathcal{X}^{s,t} - \mathcal{G}^{s,T} \times_i \{U_o^{(s,t,i)}\} \right\|_F^2 + \alpha \left\| \mathcal{G}^{s,T} - \mathcal{G}^{*(T-1)} \right\|_F^2$$



# Missing Data

- **Observation:**

- Many users may only give ratings for a few hotels at some specific timestamps

- **Solution:**

- $\mathcal{K}^s = \{(i, j, k): \mathcal{X}_{ijk}^s \text{ is available}\}$  is a triple-element set
- Objective function

$$\min_{\{G^s, \{U^{s,i}\}\}_{\geq 0}} \sum_{s=1}^M \sum_{(i,j,k) \in \mathcal{K}^s} \left( \mathcal{X}_{ijk}^s - (G^s \times_i \{U^{s,i}\})_{ijk} \right)^2 + \alpha (G_{ijk}^s - G_{ijk}^*)^2$$

# Experiment Set-up

- **Datasets:**
  - **Synthetic datasets**
  - **Real-world datasets**
    - Hotel Rating
    - Network Traffic Flow
    - Weather Forecast