# Spatiotemporal Activity Modeling via Hierarchical Cross-Modal Embedding


Yang Liu[1,2], Xiang Ao[1,2], Linfeng Dong[3], Chao Zhang[4], Jin Wang[5], and Qing He[1,2]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China [3]Zhejiang University, Hangzhou, China
[4]College of Computing, Georgia Tech, United States [5]Megagon Labs, Mountain View, United States

Email: {liuyang17z,aoxiang,heqing}@ict.ac.cn ; linfengdong22@zju.edu.cn ; chaozhang@gatech.edu ; jin@megagon.ai



ICDE2023
Anaheim, CA April 3 – 7, 2023


## Contribution

- We propose a novel hierarchical cross-modal representation learning method for spatiotemporal activity modeling, which can preserve high-order proximities in mobile data.
- We propose a flexible meta-graph based embedding framework named ACTOR, which can perform hierarchical embedding on graphs.
- We evaluate the effectiveness and efficiency of ACTOR on three real-world datasets.

## Paper, Code and Author Contact

Scan the QR code to get the paper and the code. Any questions and suggestions are welcomed.

➢ Paper: ➢ Repo: ➢ Twitter: ➢ Wechat:

## Dataset Description

Let $\mathcal{R} = \{r_1, \dots, r_N\}$ be a corpus of mobile data records. Each record $r_i \in \mathcal{R}$ is defined by a tuple $< t_i, l_i, W_i >$ where

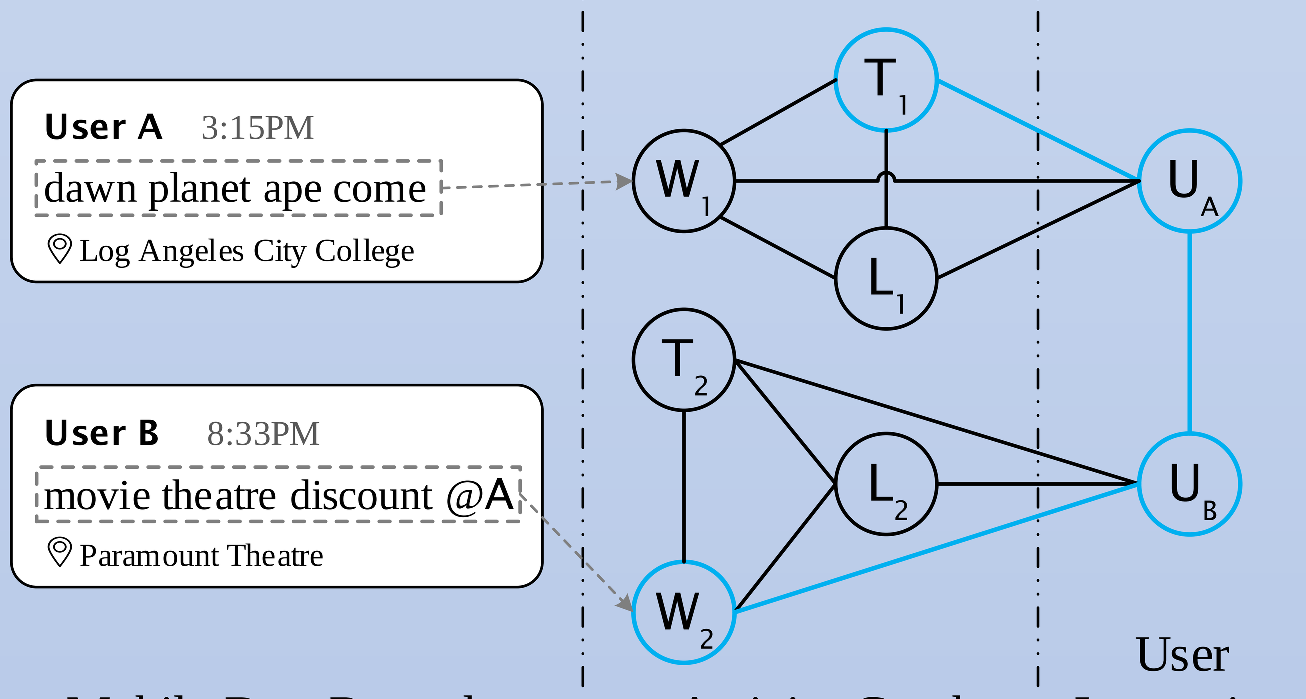
1. $t_i$ is the creating timestamp of $r_i$ ;
2. $l_i$ is a two-dimensional vector that represents the user's location when $r_i$ is created ;
3. $W_i$ is a bag of keywords denoting the text message of $r_i$ .

## Problem Definition

The problem of spatiotemporal activity modeling can be decomposed into three sub-tasks:

1. Activity prediction. Given the time, location and a text candidate set, find the most possible activity keyword.
2. Location prediction. Given the time and keywords, find the location.
3. Time prediction. Given the location and keywords, find the most possible time from a time candidate set.

## Main Frameworks



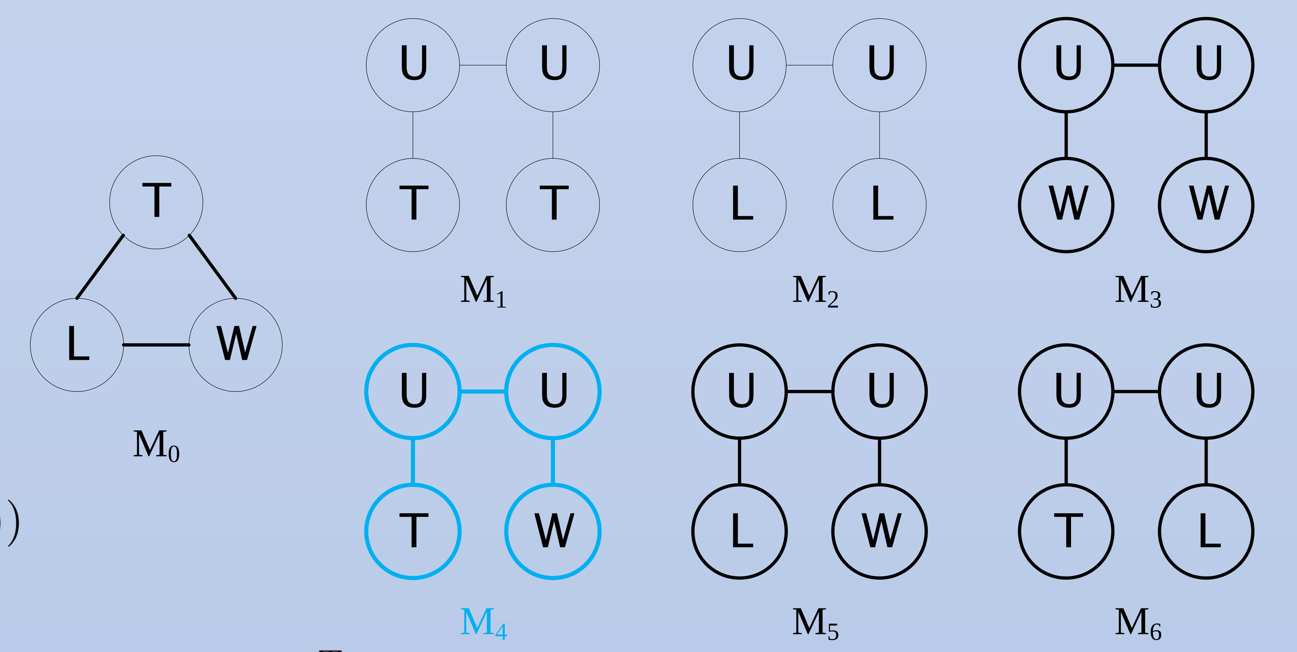

Mobile Data Records | Activity Graph | User Interaction Graph

$$p_e(v_j|v_i) = \frac{\exp(\boldsymbol{x}_j'^{\mathrm{T}} \boldsymbol{x}_i)}{\sum_{f_e(v_i,v_k)=e} \exp(\boldsymbol{x}_k'^{\mathrm{T}} \boldsymbol{x}_i)}$$

$$\hat{p}_e(v_j|v_i) = \frac{a_{ij}}{d_i^e} \qquad d_i^e = \sum_{f_e(v_i,v_k)=e} a_{ik}$$

$$J_e = \sum_{v_i \in \mathcal{V}_e} \lambda_i D(\hat{p}_e(\cdot|v_i), p_e(\cdot|v_i))$$

$$J_{\mathrm{NEG}} = -\log \sigma(\boldsymbol{x}_j'^{\mathrm{T}} \boldsymbol{x}_i) - \sum_{k=1}^{K} \mathbb{E}_{v_k \sim P(v)} \log \sigma(-\boldsymbol{x}_k'^{\mathrm{T}} \boldsymbol{x}_i)$$




## Performance: Cross-Modal Retrieval

| Data | UTGEO2011 | | | TWEET | | | 4SQ | | |
|---|---|---|---|---|---|---|---|---|---|
| Task | Text | Location | Time | Text | Location | Time | Text | Location | Time |
| LGTA | 0.3571 | 0.3440 | / | 0.4615 | 0.4439 | / | 0.5739 | 0.5409 | / |
| MGTM | 0.2993 | 0.3022 | / | 0.3615 | 0.3619 | / | 0.4538 | 0.4191 | / |
| metapath2vec | 0.5062 | 0.5267 | 0.3169 | 0.5083 | 0.5369 | 0.2986 | 0.8475 | 0.8673 | 0.3262 |
| LINE | 0.5433 | 0.5442 | 0.3427 | 0.6246 | 0.5997 | 0.3235 | 0.9076 | 0.8954 | 0.3637 |
| LINE(U) | 0.5830 | 0.5798 | 0.3578 | 0.6315 | 0.6066 | 0.3297 | 0.9078 | 0.8972 | 0.3719 |
| CrossMap | 0.5778 | 0.6015 | 0.3852 | 0.6701 | 0.6561 | 0.3439 | 0.9393 | 0.9138 | 0.3690 |
| CrossMap(U) | 0.5808 | 0.6070 | 0.3712 | 0.6894 | 0.6632 | 0.3469 | 0.9441 | 0.9137 | 0.3735 |
| ACTOR | **0.6207** | **0.6275** | **0.3885** | **0.6991** | **0.6805** | **0.3509** | **0.9519** | **0.9211** | **0.3758** |

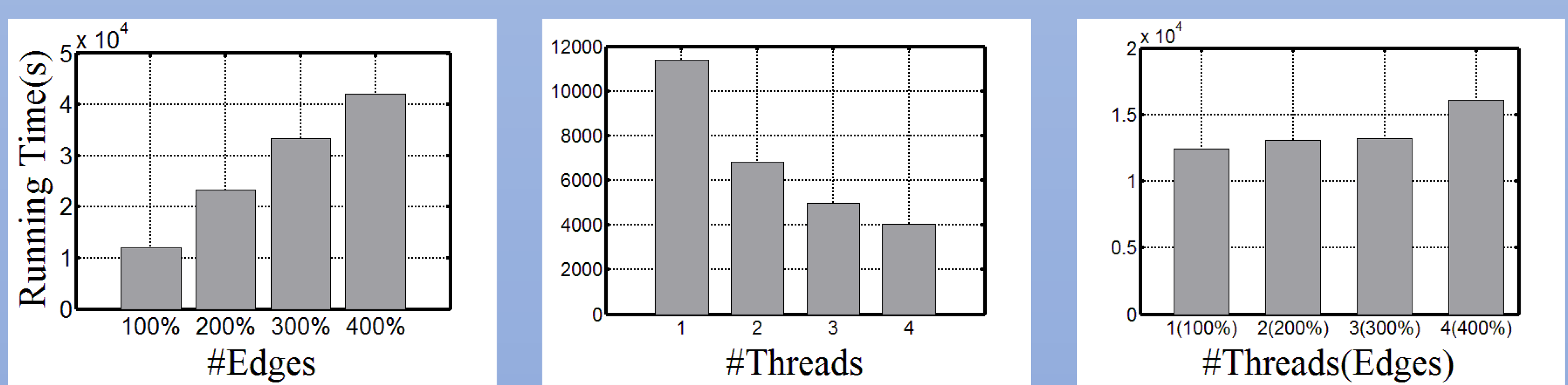
- ACTOR consistently outperform all the other methods on the three datasets, with at most 85.9 percent improvements compared with LGTA and16.0 percent improvements with CrossMap.

## Case Study


| ACTOR | | CrossMap | |
|---|---|---|---|
| Text | Time | Text | Time |
| portofla | 10:57:39 | today | 10:57:39 |
| dock | 14:34:54 | day | 17:42:27 |
| groovecruise | 17:42:27 | time | 14:34:54 |
| departure | 18:53:55 | get | 18:53:55 |
| mex | 10:13:51 | camera | 10:13:51 |
| passport | 10:38:16 | work | 13:33:17 |
| berth | 6:06:47 | another | 16:49:07 |
| ship | 16:49:07 | segundo | 15:51:17 |
| segundo | 14:59:13 | got | 10:38:16 |
| evo | 5:47:58 | hit | 14:59:13 |

- When we query the location of the port of Los Angeles, the results of ACTOR are closely related to the port, like "dock", "departure" or the place "port of LA". However, CrossMap prefers some general words like "today", "time", etc.

## Scalability




## Acknowledgement


The research work is supported by National Key R&D Plan No. 2022YFC3303303, National Natural Science Foundation of China under Grant (No.61976204). Xiang Ao is also supported by the Project of Youth Innovation Promotion Association CAS, Beijing Nova Program Z201100006820062.