Spatiotemporal-aware Session-based Recommendation with Graph Neural Networks

Yinfeng Li
Tsinghua University,
Beijing, China

Chen Gao†
Tsinghua University,
Beijing, China

Xiaoyi Du
Meituan Inc.,
Beijing, China

Huazhou Wei
Meituan Inc.,
Beijing, China

Hengliang Luo
Meituan Inc.,
Beijing, China

Depeng Jin
Tsinghua University,
Beijing, China

Yong Li
Tsinghua University,
Beijing, China

ABSTRACT

Session-based recommendation (SBR) aims to recommend items based on user behaviors in a session. For the online life service platforms, such as Meituan, both the user's location and the current time primarily cause the different patterns and intents in user behaviors. Hence, spatiotemporal context plays a significant role in the recommendation on those platforms, which motivates an important problem of spatiotemporal-aware session-based recommendation (STSBR). Since the spatiotemporal context is introduced, there are two critical challenges: 1) how to capture session-level relations of spatiotemporal context (inter-session view), and 2) how to model the complex user decision-making process at a specific location and time (intra-session view). To address them, we propose a novel solution named **STAGE** in this paper. Specifically, STAGE first constructs a *global information graph* to model the multi-level relations among all sessions, and a *session decision graph* to capture the complex user decision process for each session. STAGE then performs inter-session and intra-session embedding propagation on the constructed graphs with the proposed graph attentive convolution (GAC) to learn representations from the above two perspectives. Finally, the learned representations are combined with spatiotemporal-aware soft-attention for final recommendation. Extensive experiments on two datasets from Meituan demonstrate the superiority of STAGE over state-of-the-art methods. Further studies also verify that each component is effective.

CCS CONCEPTS

• **Information systems** → **Recommender systems**;

KEYWORDS

Spatiotemporal-aware Session-based Recommendation; Spatiotemporal Context; Graph Neural Networks

ACM Reference Format:
Yinfeng Li, Chen Gao†, Xiaoyi Du, Huazhou Wei, Hengliang Luo, Depeng Jin, and Yong Li. 2022. Spatiotemporal-aware Session-based Recommendation with Graph Neural Networks. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557458

1 INTRODUCTION

Online life service platforms such as Meituan[1] and Uber Eats[2], where users can order their favorite products or foods at any time and anywhere, have significantly changed the way people live. Distinct from the traditional e-commerce platforms, users' behaviors are highly relevant to the *spatiotemporal context* on those platforms. As illustrated in Figure 1, a user may order fast food only at workday's noon (*temporal context*) when he/she is at office (*spatial context*). Therefore, the traditional session-based recommendation [15, 20, 32], which only models the users' sequential behaviors without modeling the spatiotemporal context, cannot handle the business scenarios on the online life service platforms. Under this circumstance, we define the new research problem *spatiotemporal-aware session-based recommendation* (STSBR) which aims to predict the next interacted items in user behavior sessions given the specific spatiotemporal context.

Overall speaking, there are two closely relevant research topics, session-based recommendation and spatiotemporal activity prediction. First, existing works of traditional session-based recommendation [6, 15, 20, 28, 31, 32, 43, 44, 46] have not well explored the spatialtemporal context. Second, existing works of spatiotemporal activity prediction [2, 9, 26, 48, 52] leverage spatiotemporal context as side information to predict a single activity/behavior, ignoring the fact that users have a series of behavior (behavior session) at a specific location and a given time. In other words, users' dynamic

†Chen Gao is the corresponding author (chgao96@gmail.com).

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557458

[1] https://about.meituan.com/en
[2] https://www.ubereats.com/

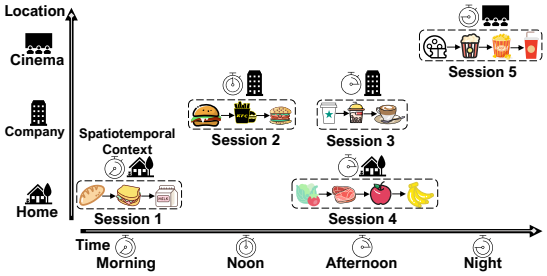

Figure 1: An illustration of spatiotemporal-aware user behavior sessions.

consumption preferences have not been well considered. In short, the above existing solutions are not suitable for the new problem of STSBR and fail to address the following two critical challenges.

- **Modeling session-level relations of spatiotemporal context is necessary but challenging (*an inter-session view*).** The traditional problem of spatiotemporal activity prediction [9, 52] only reveals the relations of different spatiotemporal contexts from the item level. In STSBR, since the behaviors in a session, rather than the single item interaction, can better represent the users' consumption intents at the specific location and time, for the session-level (an inter-session view), user preferences can be captured more accurately by extracting the relations among spatiotemporal context and sessions. However, since the relations of spatiotemporal context among distinct sessions are extremely diverse and complex, the modeling is quite challenging.
- **User decision-making process at a specifc spatiotemporal context is complex (*an intra-session view*).** In the problem of STSBR, besides the internal user intents, the spatiotemporal context also influences the user's decisions. For example, when busy at work in the office, a user may tend to launch a very short session to explore fast food for a quick lunch; while at home, the user may slowly browse a lot of foods to generate a longer session. Therefore, from the intra-session view, the behaviors in one session are largely determined by the given spatiotemporal context, which makes the decision-making process complex and thus hard to model.

To address above-mentioned challenges, we propose a method named **STAGE** (short for **S**patio**T**emporal-aware session-based recommend**A**tion with **G**raph neural n**E**tworks). Specifically, we first construct two types of graphs corresponding two challenges, *i.e.* 1) *global information graph* to represent both the item-level and session-level relations *in all sessions* and 2) *session decision graph* to accurately model the impact of both spatiotemporal context and intrinsic intents on the user's decision-making process *for each session*. To address the first challenge, we perform inter-session embedding propagation on *global information graph* to extract useful predictive signal/information from item-level and session-level relations of spatiotemporal context on the global graph. To address the second challenge, we propose to conduct intra-session embedding propagation on the *session decision graph* by first learning the user's dynamic interests in the current session and then combining the impact of both spatiotemporal context and internal intents (sub-interests) on the user's decision-making process. Finally, we combine the learned node representations from both inter- and intra-session perspectives with the proposed spatiotemporal-aware soft-attention to make final recommendation.

Table 1: Explorations of spatiotemporal context.

Model		Beijing		Shanghai	
		P@10	M@10	P@10	M@10
GRU4Rec	w/o ST	20.46	13.19	19.95	12.83
	w ST	**21.32**	**13.84**	**20.86**	**13.48**
SR-GNN	w/o ST	23.92	14.49	23.10	13.41
	w ST	**24.91**	**15.17**	**24.18**	**14.07**

In short, the main contributions of this paper are as follows.

- In this work, we take the first pioneering step to approach the problem of spatiotemporal-aware session-based recommendation, which has broad applications in real-world information systems but has not been explored by existing works.
- We conduct embedding propagation with the proposed graph attentive convolution (GAC) on the two constructed graphs, *global information graph* and *session decision graph*, for capturing the multi-level relations from the inter-session view and the complex decision process from the intra-session view.
- Extensive experiments on two real-world datasets from Meituan verify the effectiveness of our STAGE. Further studies also verify that each component of STAGE is rational.

2 MOTIVATION AND PROBLEM DEFINITION

2.1 Motivation via Empirical Explorations

In this part, we conduct experiments on two representative session-based recommendation (SBR) methods (GRU4Rec [15] and SR-GNN [44]) with two datasets collected from Meituan (the details are in section 4.1.1) to verify the effectiveness of spatiotemporal context. We compare the performances of the above methods with (w ST) and without spatiotemporal context (w/o ST). Specifically, we combine (add) the embeddings of items and spatiotemporal context (*i.e.* location and time) when generating session representations to fuse spatiotemporal information. The results are shown in Table 1. From the results, we can observe that all methods achieve significant performance gain when spatiotemporal context is added, which verifies the effectiveness of spatiotemporal context to enhance SBR in Meituan. However, the spatiotemporal context has not been well explored by existing SBR methods.

2.2 Problem Definition

Motivated by the results of the above analysis, we formally define a new problem, the spatiotemporal-aware session-based recommendation (STSBR)[3]. Let $\mathcal{I}, \mathcal{L}, \mathcal{T}, \mathcal{S}$ denote the set of items, locations, time-slots and sessions in all observed records $\mathcal{D}$. Let N_I, N_L, N_T, N_S be the number of items, locations, time-slots, and sessions, respectively. Generally, a session s can be represented as a list $[i_{s,1}, i_{s,2}, ..., i_{s,n}]$ with the chronological order of user interactions, where $i_{s,j} \in \mathcal{I}$ denotes the n-th clicked item in session s, where n denotes the length of session. Meanwhile, the spatiotemporal context of session s is denoted as (l_s, t_s), which indicates the session occurred on location $l_s \in \mathcal{L}$ at time-slot $t_s \in \mathcal{T}$. Given

[3]Note that the next POI recommendation [22] (regard POI as item) is completely different from our STSBR task (enhance recommendation with spatiotemporal context).

Table 2: The description of commonly used notations.

Notations	Description
$\mathcal{L}, \mathcal{T}, \mathcal{I}, \mathcal{S}$	The set of locations, time-slots, items and sessions.
l, t, i, s	Location ID, time-slots ID, item ID and session ID.
C / c_k	Intent set / intent ID.
$\mathcal{G}_g, \mathcal{G}_s$	Global information graph and session decision graph.
$\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}$	The set of node, edge, node type and edge type.
ϕ / ψ	The mapping function of node type / edge type.
$e = (b, a)$	The directed edge from source node a to target node b.
$\mathbf{h}_v$	The embedding of node v.
$\mathcal{N}_b^{\psi(e)}$	The neighbors of node b in edge type $\psi(e)$.
α / β	The learned attention weights.
$\mathbf{q}$ / $\mathbf{W}$ / $\mathbf{b}$	Attention vector / transition matrix / bias weights.

a session s and its spatiotemporal context (l_s, t_s), STSBR aims to predict the next item $i_{s,n+1}$ that user will interact with. Based on the above definition, the task of STSBR can be formulated as follows:
Input: The session sequence s with spatiotemporal context (l_s, t_s).
Output: A model to estimate the probability that the user of session s will click item i, formulated as $\hat{y}_s = f(i|[s, (l_s, t_s)])$.

3 METHODOLOGY

Figure 2 illustrates our proposed STAGE model, which consists of three parts, *i.e.* 1) graph construction, 2) inter- and intra-session embedding propagation, and 3) session generation and prediction. In this section, we will elaborate on the above three parts in details. We also explain the commonly used notations in Table 2.

3.1 Graph Construction

To address the two challenges in Section 1, we propose to construct 1) *global information graph* to extend the item- and session-level relations in all sessions, and 2) *session decision graph* to accurately model the impact of both spatiotemporal context and intrinsic intents on the user's decision-making process for each session.

3.1.1 Global Information Graph. Formally, let $\mathcal{G}_g = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R})$ be the constructed *global information graph* with node type mapping function $\phi : \mathcal{V} \mapsto \mathcal{A}$ and edge type mapping function $\psi : \mathcal{E} \mapsto \mathcal{R}$, where $\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}$ denote the node set, edge set, node type set and edge type set, respectively. Obviously, we have the node set $\mathcal{V} = \mathcal{L} \cup \mathcal{T} \cup \mathcal{I} \cup \mathcal{S}$ and the node type set $\mathcal{A} = \{\mathcal{L}, \mathcal{T}, \mathcal{I}, \mathcal{S}\}$, where $\mathcal{L}, \mathcal{T}, \mathcal{I}, \mathcal{S}$ denote the set of all locations, time slots, items and sessions in observed data $\mathcal{D}$ (Note that we use the same symbol to represent the set and the type at the same time for simplification). With these nodes, the multi-type edges can be constructed according to the distinct relations from two perspectives, item-level and session-level.

- **Item-level.** To capture the relations among spatiotemporal context and items at item-level, we construct seven types of edges as $\mathcal{R}^{item} = \{\mathcal{E}^{II}, \mathcal{E}^{LI}, \mathcal{E}^{IL}, \mathcal{E}^{TI}, \mathcal{E}^{IT}, \mathcal{E}^{LT}, \mathcal{E}^{TL}\}$, corresponding to different semantics. Specifically, the *item-item* edges $\mathcal{E}^{II}$ represent item transitions in all sessions. As Figure 2 (a) shows, an *item-item* edge $(i_2, i_1) \in \mathcal{E}^{II}$ means there exists a transition from item i_1 to i_2 in a certain session. The *location-item* edges $\mathcal{E}^{LI}$ and *item-location* edges $\mathcal{E}^{IL}$ represent location-item co-occurrence relations. If item i and location l appear in any session, we have $(i, l) \in \mathcal{E}^{IL}$ and $(l, i) \in \mathcal{E}^{LI}$. Similarly, to capture the time-item co-occurrence relations, we construct the *time-item* edges $\mathcal{E}^{TI}$ and *item-time* edges $\mathcal{E}^{IT}$. Finally, to model the relation of location and time in spatiotemporal context tuples, the *location-time* edges $\mathcal{E}^{LT}$ and *time-location* edges $\mathcal{E}^{TL}$ are constructed.
- **Session-level.** As emphasized in Section 1, a session is a behavior sequence of a specific user that reveals his/her personalized preferences, and modeling the impact of spatiotemporal context at the session-level is essential. Here we construct two types of edges at session-level as $\mathcal{R}^{session} = \{\mathcal{E}^{LS}, \mathcal{E}^{TS}\}$ to capture the complex relations between spatiotemporal context and sessions. Specifically, the *location-session* edges $\mathcal{E}^{LS}$ represent the distributions of sessions on a certain location. As Figure 2 (a) shows, a *location-session* edge $(l_1, s_1) \in \mathcal{E}^{LS}$ means that session s_1 occurs on location l_1. Similarly, we construct the *time-session* edges $\mathcal{E}^{TS}$ to capture the distributions of sessions at a certain time slot. Since a session just occurs under a specific spatiotemporal context but a spatiotemporal context may contains multiple sessions, we use the sessions under a certain spatiotemporal context here to enhance the modeling of spatiotemporal context.

With above settings, we can obtain the edge type set $\mathcal{R} = \mathcal{R}^{item} \cup \mathcal{R}^{session}$ and the edge set $\mathcal{E} = \{\mathcal{E}^{t_e} | \mathcal{E}^{t_e} \in \mathcal{R}\}$ for the *global information graph* $\mathcal{G}_g$ from both item- and session-level.

3.1.2 Session Decision Graph. Given that each item click in a session is determined by both user's intrinsic intents and spatiotemporal context, the goals of the session decision graph are 1) modeling the item-transition patterns and 2) capturing the user's decision-making process in each interaction. Since user behaviors in a session are always relevant to multiple aspects of intents (intent can be understood as sub-interest), we assume that all users have common K types of intents[4] and denote them as $C = \{c_1, c_2, ..., c_K\}$. Given a session s and its spatiotemporal context (l_s, t_s), the corresponding session decision graph can be denoted as $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s, \mathcal{A}_s, \mathcal{R}_s)$, for which we have node type mapping function $\phi_s : \mathcal{V}_s \mapsto \mathcal{A}_s$ and edge type mapping function $\psi_s : \mathcal{E}_s \mapsto \mathcal{R}_s$. Here $\mathcal{V}_s, \mathcal{E}_s, \mathcal{A}_s, \mathcal{R}_s$ denote the node set, edge set, node type set and edge type set, respectively. Obviously, we have a node set $\mathcal{V}_s = \mathcal{I}_s \cup C \cup \{l_s, t_s, s\}$ and a node-type set $\mathcal{A}_s = \{\mathcal{L}, \mathcal{T}, \mathcal{I}, \mathcal{S}, C\}$, where $\mathcal{I}_s \subseteq \mathcal{I}$ denotes the set consisting of items in session s. Then, we construct the multi-type edges from two views, 1) the item-transition patterns to capture dynamic interests and 2) the user's decision-making process in each item click for better user modeling.

- **Item-transition patterns.** To model the sequential patterns that reveals dynamic interests over pair-wise items in the given session s, following [31, 43], we add a self loop edge for each item and construct four types edges as $\mathcal{R}_s^{II} = \{\mathcal{E}_s^{in}, \mathcal{E}_s^{out}, \mathcal{E}_s^{in-out}, \mathcal{E}_s^{self}\}$. For example, an edge $e = (i_{s,k}, i_{s,j}) \in \mathcal{E}_s^{in}$ indicates there exists a transition from item $i_{s,j}$ to item $i_{s,k}$.
- **Decision-making process.** The decision-making process can be denoted as $user \rightarrow [intents, location, time] \rightarrow items$. Specifically, a user first combines the impact of spatiotemporal context and internal intents, and then he/she chooses an item to purchase. Here we attempt to infer user characteristics in given session by

[4]We first initialize the K common intents with intent embeddings, and then we aggregate the learned item features in each session s to update the feature of K intent nodes in session decision graph $\mathcal{G}_s$. In short, we introduce intent nodes to accurately model the user decision-making process and enhance the capability of $\mathcal{G}_s$.

imitating the inverse process of decision-making. As shown in Figure 2 (a), we construct *intent-item* edges $\mathcal{E}_s^{CI}$, *location-item* edges $\mathcal{E}_s^{LI}$, *time-item* edges $\mathcal{E}_s^{TI}$ by connecting all items to intents $c_k \in C$, location l_s and time t_s with directed edges. Further, we connect the intents $c_k \in C$, location l_s and time t_s with session node s to construct *session-intent* edges $\mathcal{E}_s^{SC}$, *session-location* edges $\mathcal{E}_s^{SL}$ and *session-time* edges $\mathcal{E}_s^{ST}$, respectively. In this way, we capture the inverse process of decision-making, *i.e.* $items \rightarrow [intents, location, time] \rightarrow session(user)$ with $\mathcal{G}_s$.

With above settings, we obtain the edge type set $\mathcal{R}_s = \mathcal{R}_s^{II} \cup \{\mathcal{E}_s^{CI}, \mathcal{E}_s^{LI}, \mathcal{E}_s^{TI}, \mathcal{E}_s^{SC}, \mathcal{E}_s^{SL}, \mathcal{E}_s^{ST}\}$ and the edge set $\mathcal{E}_s = \{\mathcal{E}_s^{te} | \mathcal{E}_s^{te} \in \mathcal{R}_s\}$ for session decision graph $\mathcal{G}_s$.

Embedding Initialization: Following [6, 43, 44], we represent each item/location/ time/intent ID with embedding vector to characterize the latent features. We denote item i, location l, time t and intent c as $\mathbf{h}_i, \mathbf{h}_l, \mathbf{h}_t, \mathbf{h}_c \in \mathbb{R}^d$, where d is the embedding size.

3.2 Inter-session Propagation Layer

To sufficiently capture the multi-level relations among all sessions, we propose to perform inter-session propagation on *global information graph* $\mathcal{G}_g$ with the proposed graph attentive convolution. Specifically, we first conduct **intra-relation aggregation** in each edge type to get the relation-specific node embeddings with different semantics. Then, we conduct **inter-relation aggregation** to get the optimally weighted combination of the semantic-specific node embeddings from all edge types with attention mechanism [37].

3.2.1 Intra-relation Aggregation. Different types of edges on $\mathcal{G}_g$ reveal distinct relations and semantics. For example, $\mathcal{E}^{II}$ means the relation of item transitions. Given a target node b and all its one-hop neighbors $\mathcal{N}_b$, we group the neighbors by distinct edge types to get $\mathcal{N}_b^{\psi(e)}$, which denotes the same type of source nodes connected to b with edge type $\psi(e)$. The intra-relation aggregation aims to learn the semantic information embedded in each relation $\psi(e)$ by aggregating the feature of the grouped source nodes $\mathcal{N}_b^{\psi(e)}$. Specifically, we conduct relation-specific propagation in multi-type edges from item- and session-level, *i.e.* $\mathcal{R} = \mathcal{R}^{item} \cup \mathcal{R}^{session}$. When conducting propagation on the edges from session-level relations, for each session s, we obtain its feature at l-th layer by computing the average value of item representations in s, formulated as,

$$\mathbf{h}_s^{(l)} = \text{Mean}(\{\mathbf{h}_{i_k}^{(l)} | i_k \in s\}). \tag{1}$$

Then we introduce how to conduct relation-specific propagation. For each target node $b \in \mathcal{V}$, we first perform message passing to capture the influence from its neighbors $\mathcal{N}_b^{\psi(e)}$ under relation $\psi(e)$. Due to the heterogeneity of nodes, different types of nodes have different feature spaces. Following [6, 13], we introduce the type-specific transformation matrix to project the features of different types of nodes into the same feature space. The message from meta relation based neighbors $a \in \mathcal{N}_b^{\psi(e)}$ at l-th layer is formulated as,

$$\mathbf{m}_{a,\psi(e)}^{(l)} = \mathbf{W}_{\psi(e)} \mathbf{h}_a^{(l-1)}, \quad a \in \mathcal{N}_b^{\psi(e)}, \psi(e) \in \mathcal{R}, \tag{2}$$

where $\mathbf{W}_{\psi(e)} \in \mathbb{R}^{d \times d}$ is the transformation matrix of edge type $\psi(e)$. $\mathbf{m}_{a,\psi(e)}^{(l)}$ and $\mathbf{h}_a^{(l-1)}$ are the original and projected feature of source node a, and its embeding at 0-th layer is $\mathbf{h}_a^{(0)} = \mathbf{h}_a$.

Aggregation with graph attentive convolution (GAC): Then, we aggregate the message from each type of edges $\psi(e)$ to learn relation-specific node embeddings. Given that each source node in $\mathcal{N}_b^{\psi(e)}$ play a different role and show different importance to target node b, we propose graph attentive convolution (GAC) to learn target node embedding as follows,

$$\mathbf{h}_{b,\psi(e)}^{(l)} = \sum\nolimits_{a \in \mathcal{N}_b^{\psi(e)}} \alpha_{ba}^{\psi(e)} \cdot \mathbf{m}_{a,\psi(e)}^{(l)}, \tag{3}$$

where $\mathbf{h}_{b,\psi(e)}^{(l)}$ is the embedding of target node b in edge type $\psi(e)$ at l-th layer. $\alpha_{ba}^{\psi(e)}$ is the attention coefficient to capture the importance of neighbor node a to target node b for edge type $\psi(e)$, which can be further computed as follows,

$$\alpha_{ba}^{\psi(e)} = \frac{\exp(\sigma(\mathbf{q}_{\psi(e)}^\top (\mathbf{h}_b^{(l-1)} \odot \mathbf{m}_a^{(l)})))}{\sum_{a' \in \mathcal{N}_b^{\psi(e)}} \exp(\sigma(\mathbf{q}_{\psi(e)}^\top (\mathbf{h}_b^{(l-1)} \odot \mathbf{m}_{a'}^{(l)})))}, \tag{4}$$

where σ denotes the activation function (we select LeakyReLU), $\odot$ denotes the *Hadamard* product, and $\mathbf{q}_{\psi(e)}$ is the attention vector for edge type $\psi(e)$.

3.2.2 Inter-relation Aggregation. After the above-introduced intra-relation aggregation, we can obtain the relation-specific node embeddings. However, the relation-specific node embeddings can only reflect information and semantics from one certain type of edge. Generally, the nodes on *global information graph* $\mathcal{G}_g$ are updated according to the messages from multi-types edges in $\mathcal{R}$. In particular, the updated nodes are in set $\mathcal{V}^* = \mathcal{L} \cup \mathcal{T} \cup \mathcal{I}$ with types $\mathcal{A}^* = \{\mathcal{L}, \mathcal{T}, \mathcal{I}\}$. The location nodes $l \in \mathcal{L}$ get relation-specific embeddings from $\mathcal{R}_\mathcal{L} = \{\mathcal{E}^{LI}, \mathcal{E}^{LT}, \mathcal{E}^{LS}\}$, the time nodes $t \in \mathcal{T}$ update embeddings via $\mathcal{R}_\mathcal{T} = \{\mathcal{E}^{TI}, \mathcal{E}^{TL}, \mathcal{E}^{TS}\}$, and item nodes $i \in \mathcal{I}$ generate relation-specific representations from $\mathcal{R}_\mathcal{I} = \{\mathcal{E}^{II}, \mathcal{E}^{IL}, \mathcal{E}^{IT}\}$. To learn more comprehensive node embeddings, we conduct inter-relation aggregation among all relation-specific embeddings by automatically learning the importance of different edge types and fusing the semantics from distinct relations with attention mechanism [37]. For each node $v \in \mathcal{V}^*$ with type $\phi(v) \in \mathcal{A}^*$, we fuse its relation-specific embeddings from edge types $\mathcal{R}_{\phi(v)}$ as follows,

$$\begin{aligned} \alpha_{v,t_e} &= \frac{\exp(\mathbf{q}_{\phi(v)}^\top \tanh(\mathbf{W}_{\phi(v)} \mathbf{h}_{v,t_e}^{(l)} + \mathbf{b}_{\phi(v)}))}{\sum_{\mathcal{E}^{t'_e} \in \mathcal{R}_{\phi(v)}} \exp(\mathbf{q}_{\phi(v)}^\top \tanh(\mathbf{W}_{\phi(v)} \mathbf{h}_{v,t'_e}^{(l)} + \mathbf{b}_{\phi(v)}))}, \\ \mathbf{h}_v^{(l)} &= \sum\nolimits_{\mathcal{E}^{t_e} \in \mathcal{R}_{\phi(v)}} \alpha_{v,t_e} \mathbf{h}_{v,t_e}^{(l)}, \end{aligned} \tag{5}$$

where $\mathbf{W}_{\phi(v)} \in \mathbb{R}^{d \times d}, \mathbf{b}_{\phi(v)} \in \mathbb{R}^d, \mathbf{q}_{\phi(v)} \in \mathbb{R}^d$ are weight matrix, bias vector and attention vector in node type $\phi(v)$, respectively. $\mathbf{h}_{v,t_e}^{(l)}$ is the relation-specific embedding of node v from edge type $\mathcal{E}^{t_e}$ at l-th layer. With the learned weights as coefficients, we can selectively aggregate information and fuse these semantic-specific embeddings to fully capture the impact of spatiotemporal context from the item- and session-level on the *global information graph*.

By stacking L propagation layers, we can capture spatiotemporal-aware information within the L-hop community of each node. We

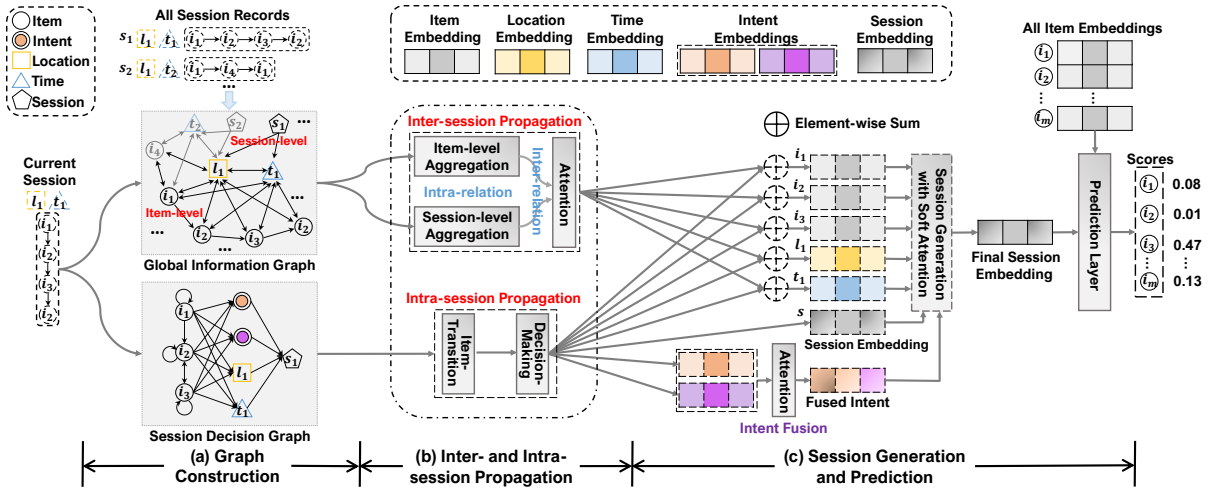

Figure 2: An overview of the proposed STAGE model. First (a), a *global information graph* is constructed with all training sessions and corresponding spatiotemporal context. Then, for each session, STAGE constructs a *session decision graph* to accurately model the decision-making process. Second (b), STAGE performs inter- and intra-session propagation on the constructed graphs to fuse information and capture the impact of spatiotemporal context. Last (c), the model combines the learned node features from two perspectives to generate session representation and further predicts the preference scores for all candidate items.

combine embeddings from all layers to obtain the final embeddings, which can be formulated as,

$$\mathbf{h}_l^g = \text{Mean}(\{\mathbf{h}_l^{(l)}|_{l=0}^{L}\}), \quad \mathbf{h}_t^g = \text{Mean}(\{\mathbf{h}_t^{(l)}|_{l=0}^{L}\}), \\ \mathbf{h}_i^g = \text{Mean}(\{\mathbf{h}_i^{(l)}|_{l=0}^{L}\}), \tag{6}$$

where $\mathbf{h}_l^g, \mathbf{h}_t^g, \mathbf{h}_i^g$ denote the final embedings of location l, time t and item i from the *global information graph* $\mathcal{G}_g$, respectively.

3.3 Intra-session Propagation Layer

The intra-session propagation on each *session decision graph* $\mathcal{G}_s$ is based on two key considerations, *i.e.* item-transition patterns and decision-making process. Specifically, we first perform **item-transition propagation** to capture the user's core and dynamic interests from item transitions. Then, we conduct **decision-making propagation** to infer user characteristics in the current session (at a certain spatiotemporal context) by imitating the inverse process of the user's decision-making process.

3.3.1 Item-transition Propagation. The user may inadvertently click on items not interested in but recorded in a session. In other words, the session sequence contains both user interests and noisy signals. To activate the user's core interests and filter out noises, we first perform the item-transition propagation to fuse interests on $\mathcal{G}_s$. Since the neighbors of the item have distinct importance to it, we utilize the proposed graph attentive convolution (GAC) to re-weight the importance of neighbors, which combine the item similarity and the impact of the decision factors.

First, we assume that each item interaction in the session is influenced by multiple decision factors (*i.e., intents, location and time*). To simplify the calculation, we generate the representation of the critical factor by computing the average value of all intents and spatiotemporal context, denoted as follows,

$$\mathbf{h}_{d_s} = \mathbf{W}_d \cdot \text{Mean}(\{\mathbf{h}_{c_k}|c_k \in C\} \cup \{\mathbf{h}_{l_s}, \mathbf{h}_{t_s}\}), \tag{7}$$

where $\mathbf{W}_d$ is a transformation matrix, and $\mathbf{h}_{c_k}, \mathbf{h}_{l_s}, \mathbf{h}_{t_s}$ denote the embeddings of intent c_k, location l_s and time t_s, respectively.

Then, we calculate the attention score of the source node i_k to target node i_j by the correlation among critical decision factor, the source node and target node, denoted as,

$$e_{jk} = \sigma(\mathbf{q}_{\psi_s(e)}^\top(\mathbf{h}_{d_s} \odot \mathbf{h}_{i_j} + \mathbf{h}_{d_s} \odot \mathbf{h}_{i_k} + \mathbf{h}_{i_j} \odot \mathbf{h}_{i_k})), \\ \alpha_{jk} = \frac{\exp(e_{jk})}{\sum_{i_{k'} \in \mathcal{N}_{i_j}^s} \exp(e_{jk'})}, \quad \psi_s(e) \in \mathcal{R}_s^{II}, \tag{8}$$

where σ denotes the activation function (LeakyReLU), $\odot$ is the *Hadamard* product, and $\mathbf{q}_{\psi_s(e)}$ is the attention vector of edge type $\psi_s(e)$. Note that $\mathcal{N}_{i_j}^s$ contains the neighbors of item i_j via all the four types of item-transition edges, *i.e.* $\mathcal{R}_s^{II} = \{\mathcal{E}_s^{in}, \mathcal{E}_s^{out}, \mathcal{E}_s^{in-out}, \mathcal{E}_s^{self}\}$.

Next, with the attention coefficients, we obtain the final output embeddings for each item in session s, which is formulated as,

$$\mathbf{h}_{i_j}^s = \sum_{i_k \in \mathcal{N}_{i_j}^s} \alpha_{jk} \mathbf{h}_{i_k}, \tag{9}$$

where $\mathbf{h}_{i_j}^s$ denotes the final embedding of item i_j from $\mathcal{G}_s$. With item-transition propagation, we filter out noises and generate item embeddings containing core and dynamic interests.

3.3.2 Decision-making Propagation. After obtaining item representations with core interests in item-transition propagation, we then attempt to infer the user characteristics in a given session at a certain spatiotemporal context via the inverse process of decision-making (*i.e.* $items \rightarrow [intents, location, time] \rightarrow session(user)$). Specifically, we first generate the feature of decision factors (*i.e.* intents, location and time) from the learned item representation. Then, we attentively aggregate the decision factors to obtain the final session embedding that contains user characteristics.

1) Generating Decision Factors. Given that the impact of decision factors on each item is different and each item contributes differently to the features of decision factors, we conduct the proposed

graph attentive convolution (GAC) to learn the embeddings of decision factors. We denote the set of decision factors as $\mathcal{F} = C \cup \{l_s, t_s\}$, where C, l_s, t_s denote intents set, location and time-slot in session s. For each decision factor $v \in \mathcal{F}$, we generate its feature as follows,

$$\begin{aligned} \alpha_{vi} &= \frac{\exp(\sigma(\mathbf{q}^\top_{\phi_s(v)}(\mathbf{h}_v \odot (\mathbf{W}_{\phi_s(v)}\mathbf{h}^s_i))))}{\sum_{i' \in \mathcal{I}_s}\exp(\sigma(\mathbf{q}^\top_{\phi_s(v)}(\mathbf{h}_v \odot (\mathbf{W}_{\phi_s(v)}\mathbf{h}^s_{i'}))))}, \\ \mathbf{h}^s_v &= \sum\nolimits_{i \in \mathcal{I}_s} \alpha_{vi}\mathbf{W}_{\phi_s(v)}\mathbf{h}^s_i + \mathbf{h}_v \end{aligned} \tag{10}$$

where σ is the activation function (LeakyReLU), $\odot$ denotes the *Hadamard* product. $\mathbf{q}_{\phi_s(v)}, \mathbf{W}_{\phi_s(v)}$ denote the attention vector and transition matrix for node type $\phi(v)$. In this way, we obtain the final decision factors $\mathbf{h}^s_v$, which consists of three types of features (intent feature $\mathbf{h}^s_{c_k}$, location feature $\mathbf{h}^s_{l_s}$ and time feature $\mathbf{h}^s_{t_s}$).

2) Generating User Characteristics[5]**.** Similarly, we further attentively aggregate the features of decision factors on $\mathcal{G}_s$ with the proposed GAC to generate final session embedding that contains user characteristics, which can be formulated as follows,

$$\begin{aligned} \alpha_{sv} &= \frac{\exp(\sigma(\mathbf{q}^\top_s(\mathbf{h}_s \odot (\mathbf{W}_s\mathbf{h}^s_v))))}{\sum_{v' \in \mathcal{F}}\exp(\sigma(\mathbf{q}^\top_s(\mathbf{h}_s \odot (\mathbf{W}_s\mathbf{h}^s_{v'}))))}, \\ \mathbf{h}^s_s &= \sum\nolimits_{v \in \mathcal{F}} \alpha_{sv}\mathbf{W}_s\mathbf{h}^s_v + \mathbf{h}_s \end{aligned} \tag{11}$$

where σ is the activation function (LeakyReLU), $\odot$ denotes the *Hadamard* product. $\mathbf{q}_s, \mathbf{W}_s$ denote the attention vector and transition matrix. The original session feature $\mathbf{h}_s$ is calculated with mean pooling according to Eq. (1). In this way, we obtain the session representation $\mathbf{h}^s_s$ on the *session decision graph* $\mathcal{G}_s$.

Following the existing SBR methods [6, 43, 44], we just perform the intra-session propagation on *session decision graph* $\mathcal{G}_s$ for one time to avoid breaking the item-transitions patterns.

3.4 Session Generation and Prediction

After the inter-session and intra-session propagation on $\mathcal{G}_g$ and $\mathcal{G}_s$, we obtain the node embeddings from both inter- and intra-session perspectives. Then we obtain the final representation of location l, time t and item i with element-wise sum pooling, formulated as,

$$\widetilde{\mathbf{h}}_l = \mathbf{h}^g_l + \mathbf{h}^s_l, \quad \widetilde{\mathbf{h}}_t = \mathbf{h}^g_t + \mathbf{h}^s_t, \quad \widetilde{\mathbf{h}}_i = \mathbf{h}^g_i + \mathbf{h}^s_i, \tag{12}$$

where $\widetilde{\mathbf{h}}_l, \widetilde{\mathbf{h}}_t, \widetilde{\mathbf{h}}_i$ denote the fused embeddings of location l, time t and item i, respectively. Given a session $s = [i_{s,1}, i_{s,2}, ..., i_{s,n}]$ and corresponding spatiotemporal context (l_s, t_s), the learned representations in session s are $[\widetilde{\mathbf{h}}_{i_{s,1}}, \widetilde{\mathbf{h}}_{i_{s,2}}, ..., \widetilde{\mathbf{h}}_{i_{s,n}}]$, $\widetilde{\mathbf{h}}_{l_s}$ and $\widetilde{\mathbf{h}}_{t_s}$. Following [43], we combine the reversed position embeddings to generate position-aware item embeddings, formulated as,

$$\mathbf{h}^*_{i_{s,j}} = \tanh(\mathbf{W}_p[\widetilde{\mathbf{h}}_{i_{s,j}} \| \mathbf{p}_{n-j+1}] + \mathbf{b}_p), \tag{13}$$

where $\mathbf{h}^*_{i_{s,j}}, \mathbf{p}_{n-j+1}$ denote the generated position-aware embedding and reversed position embedding of item $i_{s,j}$, respectively. $\mathbf{W}_p \in \mathbb{R}^{d \times 2d}$ and $\mathbf{b}_p \in \mathbb{R}^d$ are learnable parameters.

1) Item Weighting. Distinct from the previous works [31, 44] that only focus on the item similarity, we pay more attention to the user's decision-making process when assigning item weights. Specifically, we assume that the contribution of different items to the next prediction is not equal and is determined by multiple factors (i.e., intents, spatiotemporal context, and user characteristics). Hence, we first fuse the intents and then combine the above factors with spatiotemporal-aware soft-attention [47] to get item weights,

$$\begin{aligned} &\mathbf{h}^{s*}_c = \text{Attention}(\{\mathbf{h}^s_{c_k} | c_k \in C\}), \\ &\beta_j = \mathbf{q}^\top\sigma([\mathbf{W}_1\mathbf{h}^{s*}_c + \mathbf{W}_2\widetilde{\mathbf{h}}_{l_s} + \mathbf{W}_3\widetilde{\mathbf{h}}_{t_s} + \mathbf{W}_4\mathbf{h}^s_s] \odot \mathbf{h}^*_{i_{s,j}}), \end{aligned} \tag{14}$$

where $\mathbf{h}^s_{c_k}$ is the intent feature from $\mathcal{G}_s$ and we fuse all those intents with attention in Eq. (5) to get fused intent $\mathbf{h}^{s*}_c$. $\mathbf{h}^s_s$ is the local session embedding from $\mathcal{G}_s$, which contains user characteristic information. $\mathbf{W}_i|^4_{i=1} \in \mathbb{R}^{d \times d}$ and $\mathbf{q} \in \mathbb{R}^d$ are learnable parameters.

2) Session Generation. Then, we obtain the session representation by combining the item features with distinct weights, denoted as,

$$\mathbf{h}^*_s = \sum\nolimits_{j=1}^{n} \beta_j \cdot \mathbf{h}^*_{i_{s,j}}, \tag{15}$$

3) Prediction. With the obtained representation of current session $\mathbf{h}^*_s$, we first calculate the preference score of current session s on candidate item $i_j \in \mathcal{I}$ as $p_{sj} = \mathbf{h}^{*\top}_s\mathbf{h}_{i_j}$, where $\mathbf{h}_{i_j}$ is the embedding of item i_j. Let $p_s = [p_{s1}, p_{s2}, ..., p_{sN_I}]$ denotes the predicted preference scores of session s on all of the N_I items in item set $\mathcal{I}$, the recommendation probability for all candidate items can be generated with Softmax function as,

$$\hat{y}_s = \text{Softmax}(p_s) \tag{16}$$

where $\hat{y}_s = [\hat{y}_{s1}, \hat{y}_{s2}, ..., \hat{y}_{sN_I}]$ denotes the recommendation scores for all candidate items in $\mathcal{I}$.

3.5 Model Optimization

We apply the cross-entropy loss to optimize the model parameters, which is commonly used in recommendation [43, 44, 46]. For each session s, the loss function is defined as the cross-entropy of the prediction and the ground truth, formulated as follows,

$$\mathcal{L}(\hat{y}_s) = -\sum\nolimits_{i=1}^{N_I} y_{si}\log(\hat{y}_{si}) + (1 - y_{si})\log(1 - \hat{y}_{si}), \tag{17}$$

where y_s denotes the one-hot encoding vector of the ground truth in the current session.

3.6 Efficient Implementation

Given that the session number is huge in large scale datasets, the inter-session propagation on *global information graph* $\mathcal{G}_g$ (especially for the propagation on the edges from session-level) is time-consuming. Following [14], we propose an efficient sampling strategy for STAGE and further analyze the computational complexity during both the training and inference phase.

(1) Training: Given B sessions in a mini-batch, we suppose there are M_l location nodes, M_t time nodes and M_i item nodes. Since we just need the embeddings of above $M_l + M_t + M_i$ nodes, it is no need to update all the nodes in $\mathcal{G}_g$. Hence we propose neighbor sampling for the nodes that we need to update. For the inter-session propagation on G_g, we set the sampling number as N. In the item-level aggregation, the time cost for L layers' propagation is $O((M_l + M_t + M_i)N^L)$. At the session level, we update the location and time nodes with just sample sessions in the current mini-batch, which cost $O((M_l + M_t)N^L)$. For the intra-session propagation in *session decision graph*, the time cost is almost the same as existing methods

[5]The session is the sequence of behaviors made by a certain user at some time period, which reflects the characteristics of the user to some extent. Hence, we learn the session embedding to represent the user characteristics.

Table 3: Comparison with typical GNN-based SBR models.

Methods	SR-GNN	GCE-GNN	DHCN	SERec	STAGE
Spatiotemporal	no	no	no	low	high
Inter-session	×	√	√	√	√
Session-level	×	×	√	×	√
Decision-making	×	×	×	×	√

(*i.e.* GNNs, CNNs and RNNs-based), which is $O(B\overline{|S|})$, where $\overline{|S|}$ is the average session length. Hence the total time cost of STAGE in the training phase is $O((2M_l + 2M_t + M_i)N^L + B\overline{|S|})$. Compare with the existing session-based recommendation methods without spatiotemporal modeling, the additional time cost is $O((2M_l+2M_t+M_i)N^L)$, which is controllable and acceptable with appropriate sampling number N and stacking layer number L.

(2) Inference: We can only perform inter-session propagation on $\mathcal{G}_g$ in the training phase and store the learned embeddings from inter-session view. In the inference phase, we can directly look up the embeddings from inter-session view and just perform intra-session propagation on the *session decision graph* $\mathcal{G}_s$ to accurately generate representations from intra-session view. Hence, the time cost can be reduced to $O(B\overline{|S|})$, which is almost the same as of the most efficient session-based recommendation methods.

3.7 Comparison with Existing SBR models

The traditional GNN-based SBR methods only try to model the possible relations among items. In this paper, we model the impact of spatiotemporal context from both inter-and intra-session views. In the inter-session view, we construct *global information graph* at both session- and item-level. In the intra-session view, we model the user decision-making process in each session to get promising performance. SR-GNN [44] only model the item transitions from the intra-session view. Most GNN-based SBR models with inter-session modeling only model the item relations and can not capture the impact of spatiotemporal contexts, such as GCE-GNN [43] and DHCN [46]. SERec [6] can only capture the weak impact of spatiotemporal context from item-level with knowledge graph and ignores the relations at session-level. We summarize the differences between our STAGE and several typical SBR models in Table 3.

4 EXPERIMENT

4.1 Experimental Settings

4.1.1 Datasets. We conduct experiments on two real-world datasets[6] from Meituan in Beijing and Shanghai, during the time-span of Jan. 1st, 2021 - Jan. 8th, 2021. Following the definition of spatiotemporal context on Meituan business, there are 13 locations (*i.e.* "home" and "company") and 96 time-slots (dividing a day into 48 time-slots at half-hour intervals and distinguishing weekends and weekdays). Each session belongs to a certain time slot. For a fair comparison, we conduct the same preprocessing step as [43, 44, 46]. Specifically, we drop sessions of length one and items appearing less than five times, and generate sessions and corresponding labels by splitting the input session sequence for data augmentation. Furthermore, we split the first six days' data for training, the following one-day data for validation, and the last one day's data for testing. The datasets statistics are summarized in Table 4.

[6]we collected the Meituan datasets because no public datasets are suitable for our task.

Table 4: Statistics of Two Datasets from Meituan.

Dataset	#Records	#Sessions	#Items	#Loctions	#Time-slots	Avg. len.
Beijing	3,267,355	523,713	42,902	13	96	6.24
Shanghai	2,825,732	472,674	37,682	13	96	5.98

4.1.2 Evaluation Metrics. Following [20, 43, 44, 46], we adopt two widely used ranking-based metrics, **P@K** (Precision) and **M@K** (Mean Reciprocal Rank), to evaluate the model performance.

4.1.3 Compared Baselines. We compare our STAGE[7] with four categories of baseline methods, i.e., (1) classical methods that utilize neither spatiotemporal context nor session modeling (ItemKNN [33] and FPMC [32]); (2) the spatiotemporal activity prediction models (CrossMap [50], SA-GCN [48] and HAN [41]); (3) session-based recommendation methods with only intra-session modeling (GRU4Rec [15], NARM [20], STAMP [28] and SR-GNN [44]); (4) session-based recommenders with both inter- and intra-session modeling (SERec [6], GCE-GNN [43] and DHCN [46]).

4.1.4 Hyper-parameters Settings. For all the models, following [43, 44, 46], the dimension of latent vectors (*i,e.* the embedding size of items, locations, and time) is fixed as 100, the batch size is set to 100, and L_2 penalty is set as 10^{-5}. We optimize all these models with Adam [18] optimizer (the initial learning rate is 0.001 and will decay by 0.1 every three epoch). For the baselines, the hyper-parameters are initialized as the original papers and are carefully tuned to get optimal performance. Following [43], we sample 12 neighbors per node for our STAGE model. Section 4.4 reports the impact of other essential hyper-parameters in STAGE (*i.e.* model depth L and intent number K), and we use the best parameter settings in Section 4.2 and 4.3. For all methods, we run the experiment ten times and report the average results.

4.2 Overall Performance

The main results and comparisons on two datasets are shown in Table 5. From the results, we have the following observations.

- **Our proposed STAGE achieves the best performance.** Owing to the *global information graph* and *session decision graph*, STAGE can capture user's spatiotemporal-aware dynamic preferences from both inter- and intra-session views. Compared with all baselines, STAGE obtains the best performance on both datasets.
- **Session-based recommenders outperform the classical and spatiotemporal models.** Compared with classical models and spatiotemporal models, the better performance of session-based recommendation methods verifies the necessity of modeling users' dynamic interests. Among the session-based recommenders, graph-based models (*i.e.* GCE-GNN, DHCN) outperform the sequential models (*i.e.* GRU4Rec), which verifies the necessity of accurately modeling the complex relations among items and justifies our motivation to explore user's decision-making process of each item interaction with *session decision graph*.
- **The global information enhances the recommendation performance.** The inter-session methods (utilizing item transitions from all sessions) achieve the best performance among all baselines, which validates our motivation of constructing

[7]Codes are released at https://github.com/tsinghua-fib-lab/STAGE.

Table 5: Performance comparisons in % (bold* means p-value < 0.01, and bold means p-value < 0.001).**

Dataset		Beijing						Shanghai					
Category	**Method**	**P@5**	**M@5**	**P@10**	**M@10**	**P@20**	**M@20**	**P@5**	**M@5**	**P@10**	**M@10**	**P@20**	**M@20**
Classical	Item-KNN	14.16	10.28	17.65	10.94	22.44	11.32	13.71	9.97	16.72	10.57	21.83	11.03
	FPMC	13.85	10.74	16.98	11.12	22.06	11.61	13.22	10.24	15.58	10.84	21.09	11.25
Spatiotemporal (ST)	CrossMap	15.08	11.89	18.13	12.47	22.98	13.45	14.67	11.76	17.55	12.41	22.18	12.82
	SA-GCN	15.73	11.63	18.76	12.23	23.27	13.27	14.93	11.43	17.98	12.19	22.74	12.59
	HAN	16.57	12.94	19.17	13.18	23.92	14.06	15.28	12.57	18.64	12.82	23.35	13.17
Intra-Session	GRU4Rec	17.85	13.46	21.32	13.84	26.37	14.79	16.84	13.05	20.86	13.48	25.91	13.95
	NARM	18.76	13.87	23.74	14.36	29.12	15.24	17.49	13.69	22.47	13.85	28.74	14.67
	STAMP	18.33	14.02	22.83	14.85	28.89	15.43	17.37	13.76	21.94	13.92	28.48	14.81
	SR-GNN	19.48	14.27	24.91	15.17	30.25	15.98	19.15	13.53	24.18	14.07	30.23	14.99
Inter-Session	SERec	20.19	14.83	25.23	15.43	30.84	16.25	19.76	13.79	24.93	14.46	30.72	15.26
	GCE-GNN	21.07	15.24	26.04	15.83	31.69	16.43	20.16	14.25	25.41	14.95	31.36	15.48
	DHCN	20.85	15.01	25.91	15.91	31.63	16.47	20.02	14.09	25.32	14.89	31.29	15.51
ST + Inter-Session	**STAGE**	**24.50***	**18.03****	**29.99****	**18.74****	**36.20***	**19.16****	**23.34****	**16.87****	**29.20****	**17.67****	**35.86***	**18.12****

Table 6: Ablation study of inter-session propagation.

Model		Beijing		Shanghai	
		P@10	**M@10**	**P@10**	**M@10**
Inter-session	w/o $\mathcal{G}_g$	26.58	16.27	25.83	15.63
Intra-relation Aggregation	Item Level	28.95	18.33	28.14	17.08
	Session Level	28.43	18.16	27.86	16.87
	I&S Levels	**29.99**	**18.74**	**29.20**	**17.67**
Inter-relation Aggregation	Max	27.35	17.14	26.47	16.23
	Sum	29.18	18.46	28.61	17.27
	Mean	29.42	18.39	28.97	17.21
	Attention	**29.99**	**18.74**	**29.20**	**17.67**

Table 7: Ablation study of intra-session propagation.

Model		Beijing		Shanghai	
		P@10	**M@10**	**P@10**	**M@10**
Intra-session	w/o $\mathcal{G}_s$	26.37	16.13	25.65	15.51
Item-transition Propagation	Undirected	28.27	17.95	27.56	16.73
	Directed	**29.99**	**18.74**	**29.20**	**17.67**
Decision-making Propagation	w/o Intents	28.73	18.27	27.89	17.15
	w/o ST	27.62	17.37	26.84	16.51
	w/o Session	29.26	18.41	28.82	17.43
	I&ST&S	**29.99**	**18.74**	**29.20**	**17.67**

global information graph to sufficiently capture spatiotemporal information and enrich the item relations with all sessions.

4.3 Ablation Study

In this section, we performed ablation studies to evaluate the effectiveness of several key designs in our proposed STAGE.

4.3.1 Effectiveness of Inter-session Propagation. We propose to perform inter-session propagation on the *global information graph*, which contains intra-relation and inter-relation aggregation. For intra-relation aggregation, we compare the performance of the models performing propagation at the only item-level, only session-level, and both levels. For inter-relation aggregation, we compare the performance with different fusion functions (*i.e.* Max, Sum, Mean and Attention). The results in Table 6 show that STAGE suffers performance drop when removing *global information graph* (w/o $\mathcal{G}_g$), which verifies its effectiveness. Specifically, for intra-relation aggregation, the model with two levels outperforms that with the only item- or session-level. For inter-relation aggregation, the attention mechanism achieves the best performance.

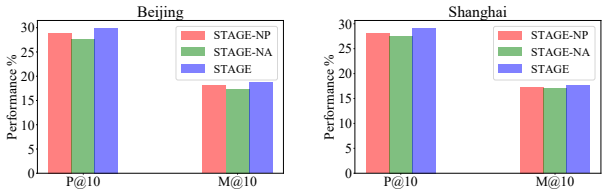

Figure 3: Impact of each component in session generation.

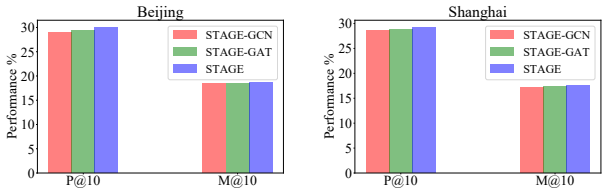

Figure 4: Performance comparison of the proposed method using different GNN aggregation schemes on two datasets.

4.3.2 Effectiveness of Intra-session Propagation. The intra-session propagation on *session decision graph*, including item-transition and decision-making propagation. For item-transition propagation, we compare the performance with undirected and directed edges for item transitions modeling. For decision-making propagation, we compare the performance of the models that without intent modeling (w/o Intents), without spatiotemporal modeling (w/o ST), without session node (w/o Session), and with above all components (I&ST&S). From the results in Table 7, we can observe that the model with directed edges achieves better performance, which means the chronological order of item transitions is crucial. As for the decision-making propagation, removing any component will lead to performance degradation. If we remove *session decision graph* (w/o $\mathcal{G}_s$), STAGE suffers significant performance drop, which verifies its effectiveness.

4.3.3 Effectiveness of Session Generation. To investigate the impact of two key designs in the session generation module, we develop two variants of STAGE, *i.e.*, the model without the reversed position embeddings (**STAGE-NP**) and without the soft attention mechanism (**STAGE-NA**). As shown in Figure 3, we can conclude that both the reversed position embeddings and soft attention are beneficial to better performance.

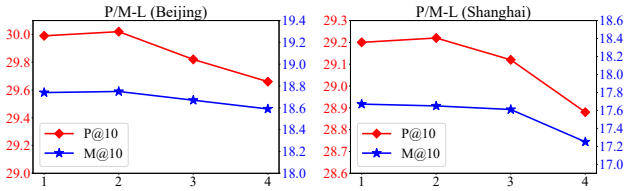

Figure 5: The impact of number of layers L on two datasets.

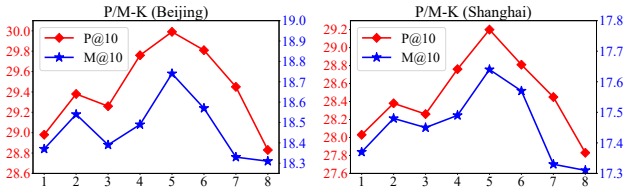

Figure 6: The impact of intent number K on two datasets.

4.3.4 Effectiveness of Graph Attentive Convolution. To evaluate the impact of the proposed graph attentive convolution (GAC), we compare the performance of models with different GNN aggregation schemes, *i.e.*, the models with GCN [19] (STAGE-GCN), with GAT [38] (STAGE-GAT) and with proposed GAC (STAGE). As shown in Figure 4, STAGE-GAT outperforms STAGE-GCN but still falls behind our STAGE model with graph attentive convolution (GAC). One possible reason is that GAC involves fewer parameters and is more efficient. This verifies the effectiveness of GAC.

4.4 Hyper-parameter Study

Impact of Model Depth L. To study the impact of model depth of *global information graph*, we vary L in $\{1, 2, 3, 4\}$. According to the results in Figure 5, STAGE achieves the best performance with one or two layers on both datasets. However, as the number of layers continues to increase, the performance drops significantly because of the over-smoothing [21]. Since the performance of the first two layers is almost equal, we set the depth L as 1 for better efficiency.

Impact of Intent Number K. To study the influence of intent number K in *session decision graph*, we vary K from 1 to 8. From the results in Figure 6, we can observe that STAGE achieves the best performance when $K = 5$ and performs the worst when $K = 1$, which verifies the rationality of modeling the user's intents on each item click. However, the performance drops when $K > 5$, which means too fine-grained intents are not conducive to the final performance. Hence, we set $K = 5$ for both datasets.

5 RELATED WORK

Session-based Recommendation. Existing works on session-based recommendation (SBR) can be categorized into three classes: classical methods, sequential-based methods, and graph-based methods. Representative classical methods include Item-KNN [33] that recommends items based on the similarity between items, and FPMC [32] that combines the matrix factorization and the first-order Markov chain for capturing both sequential effects and user preferences. Sequential-based methods [4, 17, 36] model the session as a sequence. Hidasi *et al.* propose the first work, GRU4REC, to model the session data with Gated Recurrent Unit (GRU). NARM [20] and STAMP [28] enhance GRU4Rec by incorporating attention mechanism [37] into stack GRU encoder to capture the more representative item-transition information. Graph-based methods [5, 29, 31, 40, 43, 44] propose to build a session graph to capture the complex relations among items. SR-GNN [44] is the first work to model the item transitions in a session with a directed graph and employ a gated GNN layer to obtain item embeddings. FGNN [31] designs a weighted attention graph layer for learning items embeddings. SGNN-HN [29] applies a star graph neural network (SGNN) to model the complex transition relationship between items. LESSR [5] proposes the EOPA and the SGAT layers to solve two information loss problems. GCE-GNN [43] converts the session sequences into session graphs and constructs a global graph to enhance global item transitions. DHCN [46], SHARE [40] and HIDE [24] utilize hypergraph [1, 10] to model the high-order item transitions. However, all those methods ignore the impact of spatiotemporal context, which is an important factor besides users' internal interests.

Spatiotemporal Activity Prediction. Spatiotemporal activity prediction aims to predict user activities at a specific location and time. Early studies [2, 9, 26, 49, 52] model the user's historical spatiotemporal data as tensor and conduct tensor factorization [34] to learn latent feature, such as MCTF [2]. Fan *et al.* [9] combine tensor factorization and transfer learning to recommend user's favorite apps with spatiotemporal context. Heterogeneous graph [7, 11, 16, 27, 39, 41, 45], extending the model capability with multiple types of nodes or edges, has been widely applied to recommendation [3, 8, 12, 23, 30, 35, 42, 51]. Hence, graph models [25, 48, 50] are introduced in the spatiotemporal activity prediction, which converts users, locations, time, and behaviors into a heterogeneous graph. CrossMap [50] unifies different regions, hours, and activities into a graph and learns the latent feature with graph embedding. DisenHCN [25] adopts hypergraph to obtain disentangled representations for spatiotemporal activity prediction. However, these methods only utilize static behavior information without considering the sequential patterns.

6 CONCLUSION AND FUTURE WORK

In this work, we study the new problem of spatiotemporal-aware session-based recommendation. We propose to construct two types of directed heterogeneous graphs, *i.e.* global information graph with all sessions and session decision graph for the current session, to model the complex relations among locations, time and items. We develop a GNN-based method named STAGE, which performs inter-session and intra-session propagation to obtain the embeddings that contain spatiotemporal context information and make a better recommendation. Extensive experiments on two real-world datasets collected from Meituan demonstrate the superiority of our STAGE. Further studies verify that STAGE can enhance the existing SBR methods with acceptable extra time. For future work, we plan to deploy online A/B tests to further evaluate the performance.

ACKNOWLEDGMENT

This work is supported in part by National Key Research and Development Program of China under 2020YFA0711403, by National Natural Science Foundation of China under U1936217, by the fellowship of China Postdoctoral Science Foundation under 2021TQ0027 and 2022M710006. This work is also supported by Meituan.

REFERENCES

[1] Sameer Agarwal, Kristin Branson, and Serge Belongie. 2006. Higher order learning with graphs. In *ICML*. 17–24.

[2] Preeti Bhargava, Thomas Phan, Jiayu Zhou, and Juhan Lee. 2015. Who, what, when, and where: Multi-dimensional collaborative recommendations using tensor factorization on sparse user-generated data. In *WWW*. 130–140.

[3] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle recommendation with graph convolutional networks. In *SIGIR*. 1673–1676.

[4] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential Recommendation with Graph Neural Networks. In *SIGIR*. 378–387.

[5] Tianwen Chen and Raymond Chi-Wing Wong. 2020. Handling information loss of graph neural networks for session-based recommendation. In *KDD*. 1172–1180.

[6] Tianwen Chen and Raymond Chi-Wing Wong. 2021. An Efficient and Effective Framework for Session-based Social Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 400–408.

[7] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*. 135–144.

[8] Shaohua Fan, Junxiong Zhu, Xiaotian Han, Chuan Shi, Linmei Hu, Biyu Ma, and Yongliang Li. 2019. Metapath-guided heterogeneous graph neural network for intent recommendation. In *KDD*. 2478–2486.

[9] Yali Fan, Zhen Tu, Yong Li, Xiang Chen, Hui Gao, Lin Zhang, Li Su, and Depeng Jin. 2019. Personalized Context-aware Collaborative Online Activity Prediction. *UbiComp / ISWC* 3, 4 (2019), 1–28.

[10] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3558–3565.

[11] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *CIKM*. 1797–1806.

[12] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. 2021. Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *arXiv preprint arXiv:2109.12843* (2021).

[13] William L Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. 2018. Embedding logical queries on knowledge graphs. *arXiv preprint arXiv:1806.01445* (2018).

[14] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*. 1025–1035.

[15] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[16] Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *EMNLP*. 3622–3631.

[17] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[20] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.

[21] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, Vol. 32.

[22] Yang Li, Tong Chen, Hongzhi Yin, and Zi Huang. 2021. Discovering collaborative signals for next POI recommendation with iterative Seq2Graph augmentation. *arXiv preprint arXiv:2106.15814* (2021).

[23] Yinfeng Li, Chen Gao, Xiaoyi Du, Huazhou Wei, Hengliang Luo, Depeng Jin, and Yong Li. 2022. Automatically Discovering User Consumption Intents in Meituan. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. 3259–3269.

[24] Yinfeng Li, Chen Gao, Hengliang Luo, Depeng Jin, and Yong Li. 2022. Enhancing Hypergraph Neural Networks with Intent Disentanglement for Session-Based Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 1997–2002.

[25] Yinfeng Li, Chen Gao, Quanming Yao, Tong Li, Depeng Jin, and Yong Li. 2022. DisenHCN: Disentangled Hypergraph Convolutional Networks for Spatiotemporal Activity Prediction. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE.

[26] Ziyue Li, Nurettin Dorukhan Sergin, Hao Yan, Chen Zhang, and Fugee Tsung. 2020. Tensor completion for weakly-dependent data on graph for metro passenger flow prediction. In *AAAI*, Vol. 34. 4804–4810.

[27] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *EMNLP-IJCNLP*. 4821–4830.

[28] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1831–1839.

[29] Zhiqiang Pan, Fei Cai, Wanyu Chen, Honghui Chen, and Maarten de Rijke. 2020. Star graph neural networks for session-based recommendation. In *CIKM*. 1195–1204.

[30] Yukun Ping, Chen Gao, Taichi Liu, Xiaoyi Du, Hengliang Luo, Depeng Jin, and Yong Li. 2021. User Consumption Intention Prediction in Meituan. (2021).

[31] Ruihong Qiu, Jingjing Li, Zi Huang, and Hongzhi Yin. 2019. Rethinking the item order in session-based recommendation with graph neural networks. In *CIKM*. 579–588.

[32] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[33] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[34] Amnon Shashua and Tamir Hazan. 2005. Non-negative tensor factorization with applications to statistics and computer vision. In *ICML*. 792–799.

[35] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and S Yu Philip. 2018. Heterogeneous information network embedding for recommendation. *TKDE* 31, 2 (2018), 357–370.

[36] Jing Song, Hong Shen, Zijing Ou, Junyi Zhang, Teng Xiao, and Shangsong Liang. 2019. ISLF: Interest Shift and Latent Factors Combination Model for Session-based Recommendation.. In *IJCAI*. 5765–5771.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.

[38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[39] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393* (2020).

[40] Jianling Wang, Kaize Ding, Ziwei Zhu, and James Caverlee. 2021. Session-based Recommendation with Hypergraph Attention Networks. In *SDM*. SIAM, 82–90.

[41] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The World Wide Web Conference*. 2022–2032.

[42] Yifan Wang, Suyao Tang, Yuntong Lei, Weiping Song, Sheng Wang, and Ming Zhang. 2020. DisenHAN: Disentangled Heterogeneous Graph Attention Network for Recommendation. In *CIKM*. 1605–1614.

[43] Ziyang Wang and et al. [n. d.]. Global context enhanced graph neural networks for session-based recommendation. In *SIGIR*. 169–178.

[44] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 346–353.

[45] Zeqiu Wu, Rik Koncel-Kedziorski, Mari Ostendorf, and Hannaneh Hajishirzi. 2020. Extracting Summary Knowledge Graphs from Long Documents. *arXiv preprint arXiv:2009.09162* (2020).

[46] Xin Xia, Hongzhi Yin, Junliang Yu, Qinyong Wang, Lizhen Cui, and Xiangliang Zhang. 2020. Self-supervised hypergraph convolutional networks for session-based recommendation. *arXiv preprint arXiv:2012.06852* (2020).

[47] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *arXiv preprint arXiv:1502.08029* (2015).

[48] Yue Yu, Tong Xia, Huandong Wang, Jie Feng, and Yong Li. 2020. Semantic-aware spatio-temporal app usage representation via graph convolutional network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–24.

[49] Quan Yuan, Wei Zhang, Chao Zhang, Xinhe Geng, Gao Cong, and Jiawei Han. 2017. PRED: Periodic region detection for mobility modeling of social media users. In *WSDM*. 263–272.

[50] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*. 361–370.

[51] Jun Zhang, Chen Gao, Depeng Jin, and Yong Li. 2021. Group-Buying Recommendation for Social E-Commerce. In *ICDE*. IEEE, 1536–1547.

[52] Vincent Wenchen Zheng, Bin Cao, Yu Zheng, Xing Xie, and Qiang Yang. 2010. Collaborative Filtering Meets Mobile Recommendation: A User-Centered Approach.. In *AAAI*, Vol. 10. Citeseer, 236–241.