Evaluating GNN Architecture Embeddings in Conjunction with Self-Supervised Training Objectives for the Session Disambiguation Process

Matheus H. C. Leme[1], Giovani O. Santos [1], Vinicius S. Simm[1], Douglas R. Tanno[1], Mateus Macorini Alves[1], Marcos A. Domingues[1]

[1]Department of Informatics – State Univerity of Maringá – Maringá – PR – Brazil

{pg55308, ra133166, pg404811, pg404806, pg405271, madomingues}@uem.br

Abstract. *Noisy interaction histories in recommender systems can degrade personalization, posing a structural challenge analogous to the shared accounts problem. In this work, we extract and evaluate latent representations using four GNN architectures under four self-supervised learning techniques, combined with hybridization with a sequential attention model. An empirical evaluation of 16 configurations across two domains reveals that standalone graph-based approaches suffer from representational collapse in subjective scenarios. However, the hybridization can act as an essential topological corrector, suppressing noise-induced similarities and precisely isolating the user's current intent, thereby contributing to persona disambiguation.*

1. Introduction

Session-aware recommender systems, or personalized session-based recommender systems, provide recommendations based on the analysis of two sets of information, that is, past and current records, which are present in historical sessions and the ongoing session, respectively. In this recommendation paradigm, users are not anonymized, and therefore, the wealth of information contained in the session history is preserved to model the user's long-term preferences and general behaviors, without disregarding recent interactions in the ongoing session that model short-term preferences.

Given this scenario, one of the major challenges lies in the qualification and processing of historical information. If misused, this supplementary information can degrade the final recommendation's performance rather than refine it, due to the dynamic interest observed during the ongoing session [Wang et al. 2021]. Therefore, the choice of extraction method is crucial for maximizing the potential of this information set. Although historical user records are a rich source of information capable of improving the recommendation list, in real-world scenarios, there are several phenomena that hinder this process. A significant part of the problem lies in identifying which sessions are promoters and which are detractors regarding the current session. In this regard, we propose to analyze this weakness presented by session-aware recommender systems from the perspective of the shared-account problem [Ma et al. 2019].

To exemplify the parallelism of these approaches, consider a user whose account on a music streaming platform is used on weekdays for intents aligned with the current session, but on weekends, the account is shared with family members, corresponding to sessions with divergent intents. In this case, recommendation generation will be negatively affected if the system uses the complete listening history in an agnostic manner,

due to the juxtaposition of musical tastes. In the shared-account literature, the validation of models that solve this problem is frequently conducted through synthetic data fusion, where sessions from different users are artificially mixed to create a rigorous ground truth for persona separation [Ma et al. 2019]. In our research, we adopt this exact validation premise to evaluate intent disambiguation. By treating a single user's noisy history as a composition of multiple latent intents, we simulate the severe conditions of a shared account. Thus, we subject our models to a synthetic retrieval task where the true sequential intent must be isolated from an overwhelming injection of foreign interactions, proving the robustness of the latent space in separating mixed behaviors.

In our work, we aim to mitigate this problem as efficiently as possible by isolating and removing historical sessions that are dissimilar to the ongoing session. To generate an expressive representation of these latent intents, we model sessions as graphs in which vertices and edges represent items and transitions between them, respectively. This approach captures the overall connectivity and higher-order relationships in the catalog [Wu et al. 2019]. We use Graph Neural Networks (GNNs) to extract robust representational vectors, or embeddings, of the items, which compose the intent signature of each session. However, due to the fragility of some GNNs in capturing strict sequential relationships, we consolidate a hybrid architecture. In it, the GNN embeddings serve as the foundation for a sequential attention model that aggregates dynamic dependencies. Furthermore, since we do not have explicit labels indicating the exact intent of each session, self-supervised learning is essential for training this vector space, maximizing agreement between similar sessions and mitigating historical noise.

Thus, the main contribution of our work is a critical, large-scale empirical analysis of the capabilities of GNN architectures when combined with different self-supervised training methods for the task of identifying session similarity, evaluated by the intrinsic quality of the embeddings. The experiments were conducted on two datasets covering the application domains of music and e-commerce to answer the following research questions: RQ1) What is the real impact of the choice of GNN architecture and self-supervised training objectives on the quality of embeddings for the recognition of contextualized historical sessions?; RQ2) Does the synergy between graph architectures and self-supervised methods reveal promising conditions for transfer and adoption in analogous problems, such as the shared-account problem?; and RQ3) What are the effects of hybridizing the structural embeddings of the GNN with the learning of sequential attention models? By addressing these questions, our work can provide a detailed map of the landscape and valuable insights into the inherent limitations of current practices.

2. Related Work

In this section, the related work is organized into two categories, namely shared accounts in recommender systems and embedding extraction methods. The challenge of providing personalized recommendations in shared accounts consists of modeling a mixture of behaviors originating from multiple users under the same account identifier. Historically, the literature addressed this problem by dividing it into two stages: user identification and the recommendation itself. Pioneering works employed linear subspace clustering algorithms and heuristics based on consumption time to decompose users before applying traditional methods, such as K-Nearest Neighbors or Matrix Factorization [Zhang et al. 2012]. From the perspective of unsupervised learning, notable approaches include the construction of

heterogeneous graphs and the use of random walks to differentiate preferences and group sessions by user.

Despite their initial effectiveness, two-stage approaches suffer from error propagation. To mitigate this problem, models such as π-Net and PSJNet have proposed treating the task end-to-end, using Recurrent Neural Networks combined with gating and filtering mechanisms to distinguish the information of each latent user [Ma et al. 2019], alongside multi-user identification modules guided by attention [Wen et al. 2021]. More recently, due to the limitations of recurrent networks for long sequences, the state of the art has shifted toward Graph Neural Networks equipped with attention mechanisms. Specific architectures model interactions by propagating exclusive messages for each user in the graph to disambiguate the latent identity, while extensions of these approaches incorporated the time interval between interactions and subspace alignment techniques to capture refined associations between item sequences [Zhang et al. 2025].

Recent studies employed well-established graph architectures with the explicit goal of generating high-quality embeddings. Although traditionally used for direct next-item prediction, these architectures have proven to be fundamental tools for structural interaction modeling. In our context, this ability to map complex transitions into dense vector representations is the primary requirement for evaluating whether the latent space can distinguish similar historical sessions from those that represent behavioral noise.

The industrial feasibility of embedding generation to billions of items in graphs was initially demonstrated by the integration of random walks and convolutions [Ying et al. 2018]. Subsequent research focused on optimizing and simplifying these architectures, demonstrating that heavy components of traditional formulations, such as feature transformations and non-linearities, can be safely omitted [He et al. 2020]. Instead, embedding learning exclusively through neighborhood propagation and aggregation has proven to be highly effective. Taken together, these advancements underscore the power of graph-based approaches as robust feature extractors, capable of projecting items into a space where different consumption intents can be isolated.

To further improve representational capacity in capturing specific intents, the literature combined multiple learning paradigms and graph enrichment strategies. For instance, embedding quality can be substantially improved by applying spatio-temporal contrastive learning over perturbed views of the session graph in conjunction with global collaborative signals [Wan et al. 2023]. Furthermore, user behavior modeling can be enriched by extracting latent vectors from distinct local and global graphs, subsequently fusing them through soft attention mechanisms. The direct incorporation of item features into co-occurrence graphs [Peintner et al. 2025], as well as the integration of multiple topological views encompassing local, global, and hypergraph levels, refined through contrastive learning objectives [Wang et al. 2023], emerged as effective methods for improving embedding semantics. By capturing transitions across these diverse contexts and excluding negative examples, the resulting models exhibit the robustness needed to disambiguate mixed intents.

As we can see, the literature is rich in graph architectures, contrastive training objectives, and session-based methods. However, these components are frequently proposed and evaluated in isolation with a strict focus on the recommendation metric. Thus, a sig-

nificant gap remains in the systematic and empirical investigation of the synergy between different architectures and self-supervised training objectives, particularly regarding the intrinsic quality of the representation for the task of intent identification and filtering in noisy histories, a challenge structurally analogous to shared-account disambiguation.

3. Methodology

In this section, we describe our proposed evaluation methodology. We detail the session graph construction, the employed embedding extraction architectures, the self-supervised training objectives, the fusion with the sequential attention model, and the experimental setup. Figure 1 provides an overview of the pipeline evaluated in this paper. The two investigated approaches share an identical starting point, that is, embedding extraction via the session graph. They diverge in the subsequent recommendation phase. In the first strategy (indicated by the blue arrow), structural embeddings are used directly to calculate similarities between sessions and generate a list of the most similar sessions. In the second strategy (indicated by the green arrow), these embeddings serve as the fundamental input to a sequential model, where they are refined by rigorous temporal dynamics before searching for the most similar sessions.

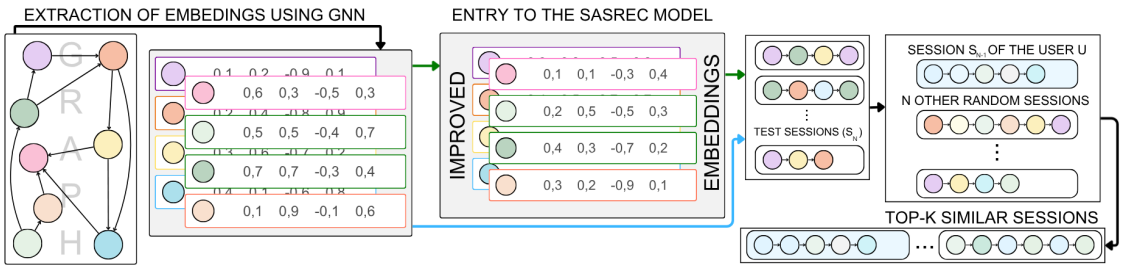


Figure 1. Overview of the proposed methodology.

3.1. Global Graph Construction and Feature Initialization

To capture the overarching behavioral patterns across the entire user base, we construct a global directed item-transition graph, denoted as $\mathcal{G} = (V, E)$. In this topology, the set of vertices V represents the unique catalog items across all datasets, rather than creating isolated graphs for individual sessions. The set of edges E captures the sequential transitions observed in the entire training corpus. Specifically, a directed edge is established between two items if they appear consecutively ($v_i \rightarrow v_{i+1}$) within a user's session. Each observed immediate transition contributes with a constant base weight. If the exact same sequential transition occurs across multiple sessions, these weights are accumulated globally. This results in a weighted adjacency matrix that directly reflects the overall frequency of strict item co-occurrences, mapping the immediate sequential transition patterns across the entire user base.

To initialize the node feature matrix $X^{(0)} \in \mathbb{R}^{|V| \times d}$, which serves as the input for the message-passing layers of the GNNs, we discard random initialization in favor of capturing prior sequential semantics. We leverage the Word2Vec framework, specifically the Skip-gram model, treating entire chronological sessions as contextual sentences. The model is trained over 100 epochs with a 5-item context window, projecting items into a fixed 128-dimensional space ($d = 128$). During the subsequent GNN optimization, this

initial dimensional space is linearly projected by the first message-passing layer to match the sampled hyperparameter dimension. This pretraining step yields an initial embedding matrix that encodes real behavioral sequences. For any isolated nodes in the global graph that lack sufficient co-occurrence data to be mapped by the Skip-gram model, we apply a Xavier uniform initialization to prevent representation collapse. Subsequently, an L_2 normalization is applied to the entire matrix $X^{(0)}$ to ensure numerical stability during the neighborhood aggregation phase of the GNN architectures.

3.2. Graph-Based Embedding Extraction Architectures

Instead of treating Graph Neural Networks (GNNs) simply as next-item predictors, our methodology repurposes them as purely structural feature extractors. The goal is to project the items into a latent space where distinct consumption intents can be geometrically isolated. To ensure a comprehensive evaluation of how different information-propagation paradigms affect this disambiguation, we selected four distinct architectures, each representing a unique approach to neighborhood aggregation.

Firstly, we employ Graph Convolutional Networks (GCN) as our isotropic baseline [Kipf and Welling 2016a]. GCNs update node embeddings through a symmetric, degree-normalized aggregation of their local neighborhoods. In the context of session histories, this architecture tests the hypothesis that a democratic smoothing of all co-occurring items is sufficient to capture the general semantic intent of a user's trajectory, without prioritizing specific transitions.

However, since noisy histories and shared accounts inherently contain divergent interactions, isotropic aggregation might blend conflicting intents. To address this, we evaluate Graph Attention Networks (GAT) [Veličković et al. 2018]. By assigning learnable attention coefficients to different neighbors, GAT performs anisotropic aggregation. This is theoretically crucial for intent disambiguation, as it allows the model to place greater weight on specific transitions that define a strong persona (e.g., a sequence of music genres) while suppressing sporadic, noisy interactions.

To investigate the impact of popularity bias and dense catalog connectivity, we incorporate GraphSAGE [Hamilton et al. 2017]. Unlike GCN and GAT, which operate over the full neighborhood, GraphSAGE samples a fixed-size subset of neighbors. In domains with highly popular items that connect to almost everything (hubs), full aggregation can lead to over-smoothing and representational collapse. GraphSAGE's inductive sampling mitigates this by restricting the receptive field, theoretically preserving the unique topological signature of niche sessions.

Finally, we test Graph Isomorphism Networks (GIN) [Xu et al. 2018], an architecture designed to be as theoretically powerful as the Weisfeiler-Lehman graph isomorphism test. For the task of separating intents, GIN is particularly relevant because it guarantees maximum discriminative capability between different graph structures. By using Multi-Layer Perceptrons (MLPs) in its update rule, GIN ensures that distinct session trajectories are mapped to distinct points in the latent space, a vital property that prevents the identities of different users from collapsing into indistinguishable vectors.

To optimize these architectures without relying on explicit session-intent labels, unavailable in the real world, we subject each GNN to four distinct Self-Supervised Learning (SSL) objectives. We evaluate DGI [Veličković et al. 2018] to test if maximizing

mutual information between local transitions and global summaries aids intent separation; GRACE [Zhu et al. 2020], a contrastive approach that evaluates the robustness of the embeddings against topological perturbations (edge dropping and feature masking); Laplacian Eigenmaps (LE) [Belkin and Niyogi 2003] to enforce strict spectral proximity for frequently co-occurring items; and VGAE [Kipf and Welling 2016b], which imposes a variational Gaussian prior over the latent space, acting as a strong spatial regularizer to forcefully separate unrelated behavioral distributions.

3.3. Hybridization with a Sequential Model

To investigate the effects of hybridizing the structural representations extracted by GNNs with explicit temporal dynamics, we adopted an architecture based on the self-attention mechanism. Specifically, we implemented an adaptation of the SASRec model, which has established itself as a reference architecture for sequential information aggregation [Kang and McAuley 2018]. The main innovation of this hybrid stage lies in the initialization of the sequential model embedding layer. Traditionally, Transformer models initialize item representations randomly. In our approach, the layer is fed with the matrix previously extracted by the best evaluated GNN paradigm configuration. This matrix is coupled to the sequential architecture without freezing its weights. This strategy allows the sequential model to inherit the rich semantic and topological similarity captured by the GNNs, while adjusting these vectors via backpropagation to capture the strict sequential transition patterns that static GNNs often overlook.

Formally, given a sequence of items consumed in a session, truncated or padded to a fixed maximum length of $L = 20$, the model processes the input by adding the pre-trained structural representation of the item to a learnable positional embedding. This positional injection is essential for the attention mechanism to become sensitive to the chronological order of the interactions, calculated by $E = E_{item} + E_{pos}$, where $E_{item}, E_{pos} \in \mathbb{R}^{L \times d}$ and $d \in \{64, 128, 256\}$ correspond to the dimensionality of the initial embeddings. The resulting matrix E is subjected to normalization and dropout layers for stabilization. Subsequently, the processed sequence feeds the Transformer encoder blocks. To maintain the autoregressive nature of the task and prevent information leakage, the model applies a causal mask to ensure that the prediction at step t depends strictly on the interactions observed up to step t. This mask is constructed using a Boolean upper-triangular matrix that zeroes out the attention weights for future positions.

The architecture was configured with 2 encoding layers, 4 parallel attention heads, and a dropout rate of 0.3 to mitigate overfitting. After propagating through the network, the hidden state corresponding to the last valid interaction in the sequence is extracted and can also be projected by a linear layer to capture the probability of each item being the next interaction in next-item recommendation tasks. Since we restrict ourselves to obtaining the embeddings learned by this model, we stop at the previous step. The evaluation protocol, detailed in Section 3.4, enables a direct comparison of the gains achieved by inserting explicit sequential modeling.

3.4. Experimental Protocol and Optimization

To empirically validate our methodology, we used two public datasets from distinct domains: music and e-commerce. Specifically, we adopted the Music4all dataset [Santana et al. 2020] (comprising 14,066 users and 98,427 tracks) and the Retailrocket

dataset [Zykov et al. 2022] (consisting of 1,404,179 users and 234,838 products, restricted strictly to view interactions). Furthermore, a rigorous preprocessing pipeline was applied to discard users with fewer than two sessions and sessions containing fewer than three items. This filtering strategy serves a dual purpose. First, it guarantees that every user contributes to both the training and testing splits (as detailed in the subsequent section); second, it ensures that each session has sufficient interaction length to accurately reflect the user's underlying intent. Without this threshold, the models would suffer from severe data sparsity, critically limiting their ability to learn robust representations [Wang et al. 2021].

After preprocessing, we obtained the final datasets. For the Music4All dataset, we retained $14,066$ users and $98,427$ items across $488,027$ sessions, with an average session length of 9.56 items and an average of 34.70 sessions per user. Regarding the Retailrocket dataset, the filtering process resulted in $12,785$ users and $58,243$ items across $41,461$ sessions, with an average session length of 7.65 items and an average of 3.24 sessions per user. We then performed hyperparameter tuning for the 16 evaluated configurations, representing all combinations of GNN architectures and self-supervised training objectives, via Bayesian optimization. Specifically, we employed the Tree-structured Parzen Estimator (TPE) algorithm implemented within the Optuna framework [Akiba et al. 2019].

3.5. Evaluation

To systematically evaluate the quality of the latent space generated by the different architectures and training objectives, we formulated an intrinsic session retrieval task. The goal of this evaluation is to verify whether the extracted embeddings are expressive enough to approximate sessions with historically aligned intents, discriminating them from a vast set of behavioral noise. Since the vectors extracted by the Graph Neural Networks represent individual items, the first step of our formulation consists of generating a unified representation for the session as a whole. To ensure geometric stability and avoid the insertion of additional weight biases in this validation step, we adopted a Mean Pooling strategy, where the latent vector of a session is defined as the average of the embeddings of all items the user interacted with in that specific session.

Once the sessions are represented in the vector space, we structured the testing protocol as a ranking problem. For each current test session S_n of a given user, we isolated the immediately preceding session S_{n-1} from the training set. This preceding session serves as our ground truth, assuming it reflects the user's most recent intent. We acknowledge an inherent tension in this formulation: in genuine shared-account scenarios, divergent personas may alternate rapidly, meaning S_{n-1} could plausibly reflect a divergent persona rather than the currently active one. However, in the absence of explicit intent labels, we adopt short-term recency as the most reliable heuristic indicator. Thus, our evaluation is specifically designed for contexts where temporal proximity strongly correlates with intent continuity. To simulate the characteristic noise of shared-account problems or diffuse histories, we constructed a candidate set containing this single correct session amalgamated with N negative samples. Following the established synthetic data fusion protocols for intent disambiguation [Ma et al. 2019], these negative sessions are randomly sampled from other users' histories ($N = 99$). This specific sampling strategy is crucial because it guarantees a rigorous zero-leakage environment where the background noise is strictly composed of foreign intents. If the model can successfully

retrieve the true chronological intent (S_{n-1}) from this highly polluted synthetic pool, it proves its capacity to filter out divergent personas, the fundamental requirement for solving the shared-account problem.

The intent retrieval process is then conducted by computing the cosine similarity between the vector of the current session S_n and all vectors in the candidate set. The candidates are sorted in descending order based on this similarity score. Basically, if the latent space is highly discriminative, the user's own S_{n-1} session should occupy the first position in the global ranking, indicating that the model understood the context transition, ignoring the intents of third parties. From this ranking and the top-K positions, we extracted the set of Information Retrieval Ranking and Latent Space Analysis metrics.

Finally, it is essential to explicitly delimit the scope of our evaluation protocol to justify the absence of traditional session-based recommendation baselines, such as SR-GNN or NARM. The primary objective of this research is not to propose a novel end-to-end architecture to achieve state-of-the-art performance on the extrinsic task of next-item prediction. Instead, we conducted an intrinsic probe into the geometric properties of the latent space with respect to intent disambiguation. Traditional recommendation baselines are heavily optimized for extrinsic ranking using supervised signals, typically without explicit spatial regularization tailored to isolate distinct personas in a zero-label environment. Consequently, comparing purely self-supervised structural embeddings against supervised end-to-end predictors in a strict cosine-similarity-based session retrieval task would introduce a structural mismatch. Therefore, our evaluation methodology focuses exclusively on isolating and comparing the foundational capabilities of different GNN paradigms and SSL objectives to geometrically filter out behavioral noise, which is the core prerequisite for addressing the shared-account phenomenon.

Given the retrieval of a single target session (S_{n-1}) amidst $N = 99$ negative samples, we employ HitRate (HR@K) that measures the proportion of cases in which the target session appears within the top-K positions (higher values indicate effective noise filtering and successful intent isolation); Mean Reciprocal Rank (MRR@K) that evaluates the exact rank position of the single ground-truth target, making it the most stringent ranking metric in our setting (an MRR close to 1.0 indicates that the model consistently prioritizes the correct intent, reducing the risk of error propagation); and Normalized Discounted Cumulative Gain (NDCG@K) that applies a logarithmic discount to lower-ranked positions, providing a sensitive assessment of relevance ordering within the top-K candidates (higher values indicate a well-structured similarity space).

To assess the geometric properties of the embedding space, we evaluate Mean Intra-User Similarity (SMIntra), which is the average cosine similarity between S_n and its ground-truth predecessor S_{n-1} (higher values indicate that the GNN preserves temporal semantic consistency across consecutive sessions of the same user); Mean Inter-User Similarity (SMInter), which is the average cosine similarity between S_n and N randomly sampled sessions from other users (lower values indicate stronger intent separation and greater embedding-space uniformity, mitigating anisotropic collapse into a shared dense region); and Confidence Gap (CG) that is the absolute difference in cosine similarity between the top-1 and top-2 ranked candidates (higher values indicate stronger confidence in intent identification, whereas lower values suggest a more ambiguous latent space that is more susceptible to false positives).

4. Results and Discussion

In this section, we present the results of our work, summarized in Table 1. The evaluation was conducted in the music and e-commerce domains, allowing us to observe how different consumption characteristics affect the architectures.

Table 1. Performance evaluation of models in the music and e-commerce domains. Bold and underlined values indicate the best metrics observed in extraction models with and without sequential information, respectively.

			Music4All						Retail Rocket					
Architecture	Training	Extractor	HR	MRR	NDCG	SMIntra	SMInter	CG	HR	MRR	NDCG	SMIntra	SMInter	CG
GCN	DGI	GNN	0.46	0.28	0.32	0.93	0.87	0.008	0.72	0.61	0.63	0.73	0.41	0.119
		SASRec	0.56	0.38	0.42	0.68	0.48	0.055	0.74	0.63	0.66	0.73	0.39	0.127
	LE	GNN	0.49	0.31	0.35	0.72	0.44	0.042	0.73	0.63	0.65	0.63	0.17	0.187
		SASRec	0.58	0.39	0.44	0.45	**0.08**	**0.093**	0.75	0.65	0.67	0.65	0.18	0.198
	GRACE	GNN	0.50	0.32	0.36	0.80	0.63	0.032	0.74	0.63	0.66	0.73	0.35	0.138
		SASRec	0.54	0.36	0.40	0.77	0.58	0.041	0.74	0.64	0.66	0.73	0.34	0.144
	VGAE	GNN	0.48	0.31	0.35	0.76	0.59	0.037	0.75	0.65	0.67	0.60	0.07	0.212
		SASRec	0.56	0.38	0.42	0.61	0.36	0.071	0.76	0.65	0.68	0.60	0.07	0.221
GAT	DGI	GNN	0.29	0.14	0.18	0.90	0.76	0.001	0.67	0.57	0.59	0.98	0.97	0.003
		SASRec	0.55	0.37	0.41	0.68	0.40	0.053	0.71	0.57	0.60	**0.82**	0.62	0.060
	LE	GNN	0.37	0.21	0.25	0.99	0.98	0.002	0.70	0.59	0.61	0.79	0.53	0.079
		SASRec	0.57	0.38	0.43	0.67	0.44	0.057	0.77	**0.66**	0.68	0.66	0.20	0.168
	GRACE	GNN	0.48	0.29	0.34	0.97	0.94	0.004	0.70	0.60	0.62	0.86	0.70	0.058
		SASRec	**0.59**	0.39	0.44	0.72	0.50	0.054	0.77	**0.66**	0.68	0.66	0.28	0.158
	VGAE	GNN	0.44	0.27	0.31	0.87	0.78	0.016	0.72	0.62	0.64	0.53	0.09	0.230
		SASRec	**0.59**	**0.40**	**0.45**	0.51	0.19	0.091	0.76	**0.66**	**0.69**	0.54	**0.06**	**0.253**
GIN	DGI	GNN	0.37	0.20	0.24	0.99	0.97	0.001	0.66	0.55	0.58	0.98	0.95	0.009
		SASRec	0.57	0.39	0.43	0.53	0.23	0.079	0.74	0.61	0.64	**0.82**	0.59	0.079
	LE	GNN	0.42	0.24	0.28	0.98	0.97	0.002	0.71	0.59	0.62	0.83	0.59	0.056
		SASRec	0.57	0.39	0.43	0.59	0.30	0.071	0.77	0.65	0.68	0.66	0.24	0.155
	GRACE	GNN	0.44	0.23	0.28	0.81	0.58	0.019	0.68	0.46	0.51	0.68	0.10	0.049
		SASRec	**0.59**	**0.40**	0.44	0.58	0.25	0.079	0.76	0.61	0.65	0.67	0.08	0.115
	VGAE	GNN	0.45	0.26	0.31	0.97	0.93	0.004	0.70	0.57	0.60	0.78	0.48	0.059
		SASRec	0.52	0.35	0.39	**0.89**	0.81	0.017	0.72	0.60	0.63	0.77	0.47	0.067
SAGE	DGI	GNN	0.46	0.29	0.33	0.89	0.82	0.015	0.73	0.61	0.64	0.66	0.24	0.160
		SASRec	**0.59**	**0.40**	**0.45**	0.63	0.35	0.073	0.77	**0.66**	0.68	0.63	0.17	0.207
	LE	GNN	0.40	0.20	0.25	0.98	0.96	0.001	0.70	0.56	0.59	0.84	0.59	0.042
		SASRec	0.58	0.39	0.43	0.67	0.41	0.059	0.77	0.65	0.68	0.71	0.30	0.125
	GRACE	GNN	0.51	0.34	0.38	0.85	0.76	0.023	0.74	0.61	0.64	0.67	0.14	0.118
		SASRec	**0.59**	**0.40**	**0.45**	0.48	0.16	0.092	**0.78**	**0.66**	**0.69**	0.64	0.07	0.180
	VGAE	GNN	0.52	0.33	0.38	0.86	0.70	0.022	0.74	0.64	0.66	0.70	0.27	0.140
		SASRec	0.56	0.38	0.42	0.82	0.67	0.033	0.75	0.65	0.67	0.71	0.27	0.149

In response to the first research question (RQ1), the ability of graph architectures to serve as isolated semantic extractors depends on the topological density of each domain. In e-commerce, the dense co-viewing characteristics favored local aggregations, with a clear standout being the GCN architecture combined with the variational objective VGAE, which achieved a Hit Rate of 0.75 and a Mean Reciprocal Rank of 0.65. On the other hand, entertainment domains presented more fluid and subjective transitions, hindering static extraction. In the musical scenario, there was no single superior isolated model; the SAGE configurations, when combined with GRACE and VGAE, achieved very similar performance, reaching Hit Rate metrics of 0.51 and 0.52, respectively, indicating that inductive neighborhood sampling handles musical co-occurrences better. This evidence shows that relying exclusively on information transmission in graphs to identify similarities between sessions is insufficient in highly subjective domains.

By synthesizing the experimental evidence, we can directly address the second research question (RQ2). We seek to evaluate whether the latent space can isolate intents and filter out the noise from diverse users, simulating the resolution of the shared-account problem. This capacity is measured by the relationship between intra-user alignment,

inter-user uniformity, and the generated confidence gap. In e-commerce, the GCN configuration, combined with VGAE, achieved exemplary disambiguation metrics: an intra-user similarity of 0.60 and an inter-user similarity of 0.07, resulting in an expressive confidence gap of 0.212. Variational inference acts as a strong spatial regularizer, forcing a normal distribution that pushes away unrelated intents. However, formulations based on mutual information, denominated DGI, failed to create useful decision boundaries, collapsing all purchase intents into inexpressive mean vectors with inter-user metrics above 0.95. This manifestation of representational collapse is a recognized vulnerability of message-passing algorithms operating on highly skewed graphs. The collapse primarily results from over-smoothing, which is severely exacerbated by popularity bias. In subjective domains, a small fraction of highly popular items act as massive topological hubs. During the recursive neighborhood aggregation phase, these hubs broadcast their dominant latent signals to almost all other nodes in the network. When combined with global optimization objectives like DGI, which forces local patches to agree with this already heavily skewed global summary, the structural distinctiveness of niche items is completely diluted. Consequently, the entire catalog is drawn toward a single dense region on the hypersphere, making it difficult to distinguish among behavioral personas.

Finally, the experimental results conclude the investigation into the effects of hybridizing structural representations with explicit temporal dynamics (RQ3). To explicitly quantify the contribution of the sequential attention module, this evaluation concurrently serves as an ablation study. By contrasting the standalone structural GNN extractors against their SASRec-hybridized counterparts, we observe that the introduction of the self-attention mechanism served as both a performance equalizer and a vital topological corrector across all domains. In the musical scenario, hybridization elevated several models to the retrieval ceiling, with configurations such as GAT combined with VGAE reaching a Hit Rate of 0.59. The most prominent result occurred in spatial restructuring, where the GCN model, combined with LE, coupled with sequential attention, adjusted its intra-user similarity to 0.45 and reduced the similarity with noise to 0.08, generating the highest confidence gap in the musical domain at 0.093. In e-commerce, a natural retrieval limit was consolidated for multiple models. The GAT architecture, combined with VGAE and self-attention, presented itself as the best option for rigorous identity filtering. By keeping the intra-user similarity at 0.54, it reduced the inter-user metric to 0.06, producing the highest confidence gap of the entire research at 0.253. Thus, we can see that hybridization is essential when facing noisy histories, because while the graph maps global relationships roughly, sequential attention fine-tunes to isolate the ongoing intent in the session.

5. Conclusion and Future Work

This work investigated the impact of different GNN and self-supervised objectives in the creation of latent spaces capable of disambiguating intents in noisy histories, addressing the inherent challenge of session-aware systems from the perspective of shared accounts. Our empirical analysis revealed that the exclusive reliance on static topologies for extracting inter-session similarities is insufficient and often dangerous. Models based on mutual information maximization repeatedly failed to separate intents into anisotropic spaces, rendering persona separation impossible, especially in entertainment domains. In contrast, variational inference proved highly effective as a spatial regularizer in e-commerce.

The main finding of this research lies in the corrective power of sequential hybridization. The injection of static structural representations into a sequential attention model proved to be prominent. The self-attention mechanism not only raised the ceiling for ranking metrics across all scenarios but also aggressively restructured the latent space, reducing similarity to other users' noise and establishing clear decision boundaries, as evidenced by the expressive confidence gaps achieved. Consequently, one can explore the refinement of historical information in scenarios of session-aware recommender systems.

While our simple mean aggregation for session representation avoids trainable biases, it equally weights all interactions, overlooking the greater semantic relevance of recent items. To address this, future work will replace it with dynamic intra-session attention mechanisms. Furthermore, since popularity bias proved a critical limitation, distorting message passing and causing dimensional collapse in non-hybrid architectures within subjective domains, we plan to construct experimental protocols fusing antagonistic profiles to evaluate the hybrid model's capacity to deterministically separate independent trajectories. Finally, we will evaluate the impact of filtering similar historical information on recommendations and incorporate rigorous statistical significance tests.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. (2020). Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Kang, W.-C. and McAuley, J. (2018). Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206. IEEE.

Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks.

Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders.

Ma, M., Ren, P., Lin, Y., Chen, Z., Ma, J., and Rijke, M. d. (2019). π-net: A parallel information-sharing network for shared-account cross-domain sequential recommendations. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 685–694.

Peintner, A., Moscati, M., Parada-Cabaleiro, E., Schedl, M., and Zangerle, E. (2025). Unsupervised graph embeddings for session-based recommendation with item features.

Santana, I. A. P., Pinhelli, F., Donini, J., Catharin, L., Mangolin, R. B., Feltrim, V. D., Domingues, M. A., et al. (2020). Music4all: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404.

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2018). Deep graph infomax.

Wan, Z., Liu, X., Wang, B., Qiu, J., Li, B., Guo, T., Chen, G., and Wang, Y. (2023). Spatio-temporal contrastive learning-enhanced gnns for session-based recommendation. *ACM Transactions on Information Systems*, 42(2):1–26.

Wang, F., Gao, X., Chen, Z., and Lyu, L. (2023). Contrastive multi-level graph neural networks for session-based recommendation. *IEEE Transactions on Multimedia*, 25:9278–9289.

Wang, S., Cao, L., Wang, Y., Sheng, Q. Z., Orgun, M. A., and Lian, D. (2021). A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7):1–38.

Wen, X., Peng, Z., Huang, S., Wang, S., and Yu, P. S. (2021). Miss: A multi-user identification network for shared-account session-aware recommendation. In *International Conference on Database Systems for Advanced Applications*, pages 228–243.

Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., and Tan, T. (2019). Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 346–353.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks?

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 974–983.

Zhang, A., Fawaz, N., Ioannidis, S., and Montanari, A. (2012). Guess who rated this movie: Identifying users through subspace clustering. *arXiv preprint arXiv:1208.1544*.

Zhang, J., Zhao, Z., Li, C., and Yu, Y. (2025). Lightweight yet fine-grained: A graph capsule convolutional network with subspace alignment for shared-account sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13242–13250.

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., and Wang, L. (2020). Deep graph contrastive representation learning.

Zykov, R., Noskov, A., and Anokhin, A. (2022). Retailrocket recommender system dataset.