Graph Representation Learning for Game Provenance

Sidney Melo[1], Aline Paes[1], Esteban Clua[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
24210-346 – Niterói – RJ – Brasil

sidneyaraujomelo@gmail.com, {alinepaes,esteban}@ic.uff.br

***Abstract. Introduction**: Game Provenance Graphs model game sessions by capturing game elements, states, and interactions, reflecting the domain's heterogeneity, dynamicity, and spatiality. Graph Neural Networks (GNNs), widely used in machine learning, learn representations through graph structures. While prior work has attempted to combine GNNs with domain-specific graph modeling, it often overlooks the inherent heterogeneity of game environments. **Objective:** The thesis aims to identify the main challenges in integrating Provenance Graphs with Representation Learning, introduce a solution that explicitly handles the heterogeneous nature of Game Provenance Graphs, and assess its effectiveness in Game Analytics tasks. **Methodology or Steps**: To reach these goals, this paper revisits core concepts of Provenance Graphs and GNN-based Representation Learning, reviews existing approaches in literature and applications in digital games, introduces the PinGLR framework tailored to game heterogeneity, and releases a novel dataset of provenance graphs. **Results:** The contributions include: (i) the PinGLR framework, which addresses heterogeneity and supports end-to-end ML pipelines, (ii) a novel method for managing node heterogeneity via intersecting attribute sets, and (iii) empirical evidence showing that PinGLR achieves comparable or better performance than traditional models in Game Analytics tasks.*

***Keywords** Representation Learning, Game Provenance Graph, Graph Embedding, Graph Neural Networks, Game Analytics.*

Resumo. Introdução**: Grafos de Proveniência de Jogos modelam sessões de jogos ao capturar elementos de jogos, seus estados e interações, refletindo a heterogeneidade, dinamicidade e espacialidade de seu domínio. Redes neurais de grafos (GNNs), amplamente usadas no aprendizado de máquina, aprendem representações vetoriais a partir de estruturas de grafos. Embora trabalhos anteriores tenham tentado combinar GNNs com a modelagem de grafos específicos a domínios, estes normalmente negligenciam a inerente heterogeneidade do domínio dos jogos. **Objetivo:** A tese tem como objetivo identificar os principais desafio na integração de Grafos de Proveniência de Jogos ao Aprendizado de Representação de grafos, introduzir uma solução que lida explicitamente a natureza heterogênea dos Grafos de Proveniência de Jogos, e avaliar sua efetividade em tarefas de Game Analytics. **Metodologia ou Passos**: Para alcançar esses objetivos, a tese revisita os conceitos principais de Grafos de Proveniência e Aprendizado de Representação baseado em GNN, revisa abordagens existentes na literatura e aplicações em jogos* *digitais, introduz o framework PinGLR, e apresenta um novo dataset de grafos de proveniência.* ***Resultados: *As contribuições incluem: (i) o framework PinGLR, que aborda a heterogeneidade e oferece suporte a pipelines completos de aprendizado de máquina, (ii) um método inovador para lidar com a heterogeneidade dos nós por meio da interseção recorrente de conjuntos de atributos, e (iii) evidências empíricas demonstrando que o PinGLR alcança desempenho comparável ou superior ao de modelos tradicionais em tarefas de Game Analytics.*

Palavras-Chave *Aprendizado de Representações, Grafo de Proveniência de Jogos, Embeddings de Grafos, Redes Neurais de Grafos, Game Analytics.*

1. Introduction

Representation learning enables automatic feature extraction from raw data and has become essential in machine learning, especially for graph-structured data, which models complex relationships across diverse domains [Bengio et al. 2013, Le-Khac et al. 2020, Myers et al. 2014]. In games, Game Provenance Graphs (GPGs)—introduced via the PinG framework—structure gameplay data as graphs of entities, interactions, and attributes, serving as a valuable foundation for Game Analytics [Kohwalter et al. 2012, Drachen et al. 2013, Ke et al. 2022]. However, existing methods like PingUMiL [Melo et al. 2019, Melo et al. 2020] fail to handle aspects present in GPGs. In the case of PingUMIL, padding strategies that limit learning quality are employed to deal with heterogeneity. Motivated by this gap, the thesis identifies heterogeneity, dynamicity, and spatiality as essential aspects to address, with heterogeneity being the most critical due to the wide variability in game elements.

Finally, the main goal of the thesis is to propose PinGLR, a novel Graph Representation Learning framework designed for Game Provenance Graphs (GPGs) that leverages their heterogeneity to enable end-to-end solutions for Game Analytics and Game Data Science. To validate PinGLR, the work explores four hypotheses: (H1) tailoring GNNs to game-specific aspects improves learning on GPGs; (H2) leveraging feature set relationships enhances heterogeneous projections; (H3) end-to-end graph approaches remove the need for manual feature engineering; and (H4) such approaches can match or outperform traditional metric-based methods. Our contributions include (i)a literature review on GPGs and Graph Representation Learning; (ii) creation of a dataset of 69 GPGs from a multiplayer airplane battle game; (iii) analysis of HGNN limitations; (iv) design and evaluation of three feature set-based projection methods; (v) proposal of the PinGLR framework; and (vi) empirical comparison of PinGLR with traditional Game Analytics solutions.

2. Background

This chapter briefly overviews fundamental topics: Provenance, Game Provenance Graphs and Graph representation learning.

Provenance, traditionally used to document the history of art or digital objects [Kohwalter et al. 2012, Moreau et al. 2011], was adapted to games by Kohwalter et al. through the PinG framework [Kohwalter et al. 2012], which maps game elements to provenance graph nodes: players and NPCs as Agents, items as Entities, and actions as Activities, with causal relationships forming the edges. This was later implemented as PinGU, a Unity plugin enabling low-effort, domain-independent extraction of Game Provenance Graphs [Kohwalter et al. 2017, Kohwalter et al. 2018].

Game Provenance Graphs enable the extraction of game metrics, pattern discovery, maintenance support, and guidance for future development [Melo 2019]. Various disciplines contribute to these goals, including data visualization [Kohwalter et al. 2016], exploratory data analysis [Alencar 2020], logical reasoning [Figueira et al. 2018], and representation learning [Melo et al. 2020]. In this work, we highlight specific characteristics of Games and how they relate to Game Provenance Graphs, which we call aspects from now on.

Games are complex and highly variable, even within the same genre, featuring diverse elements with different attributes and behaviors, making them inherently *heterogeneous*. This heterogeneity appears at the attribute level and through overlapping attribute sets across node types. Games are also *dynamic*, as game elements change state over time, and this temporal evolution can be captured through time-based subgraphs (snapshots) of Game Provenance Graphs (GPGs), even though provenance graphs are static by definition. Finally, while games exhibit *spatiality*, implicit spatial influences (e.g., a player reacting to unseen threats) are not easily captured by standard provenance recording. Therefore, GPGs capture the heterogeneous, dynamic, and spatial nature of games—characteristics that must be addressed for accurate metric extraction, pattern detection, and, in our case, representation learning.

Graphs are widely used to model relationships between entities [Hamilton et al. 2017, Angles e Gutierrez 2008], making machine learning on graphs increasingly important. Early methods depended on hand-crafted features, which were limited and labor-intensive [Hamilton et al. 2017]. Graph representation learning addressed this by embedding nodes or subgraphs into low-dimensional vectors suitable for tasks like classification and clustering [Chen et al. 2020]. A key technique in this area is Graph Neural Networks (GNNs), which use Neural Message Passing (NMP) to iteratively update node representations via message exchange [Gilmer et al. 2017]. GNNs have since evolved to handle different graph types, including Heterogeneous GNNs (HGNNs) [Zhang et al. 2019, Cen et al. 2019, Wang et al. 2020] and Dynamic GNNs (DGNNs) [Kazemi et al. 2020, Skarding et al. 2021]. This adaptability motivates the design of PinGLR, a framework aligning GPG characteristics with suitable GNN architectures.

3. PinGLR framework

We propose a novel in this chapter a novel framework for task-agnostic machine learning applications on Game Provenance Graphs based on Graph Representation Learning, which we call the PinGLR[1] (Provenance in Games representation LeaRning) framework and understand as the main contribution of this research.

By following the framework outline presented in Figure 1, one can propose and implement GNN architectures for GPG-based Game Analytics tasks. The PinGLR framework consists of two main steps and three components. The first step focuses on

[1] https://github.com/sidneyaraujomelo/PinGLR

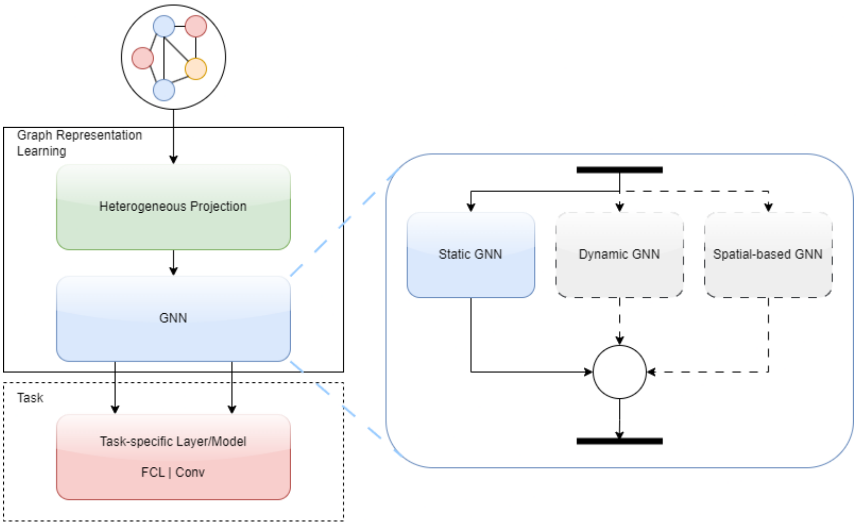

Figure 1. PinGLR conceptual framework.

graph representation learning and includes: (1) a Heterogeneous Projection component that maps heterogeneous Game Provenance Graph (GPG) nodes into a shared vector space by leveraging attribute intersections, and (2) a GNN, which learns node embeddings based on graph structure. Extensions for dynamicity and spatiality can run in parallel GNNs and be combined via mechanisms like pooling or attention. The second step is the downstream task, handled by a task-specific layer or model (e.g., a classifier or link predictor). This step is optional and allows for either supervised or self-supervised usage of the learned embeddings.

4. Game Provenance Graph Heterogeneity

We propose three mechanisms for leveraging feature set information while mapping heterogeneous node input vectors into a single-dimensional space. Most available works assume that heterogeneous node input vectors are independently distributed and that the feature sets that compose such vectors are disjoint. However, as heterogeneous graphs are used to model more complex domains, grouping node feature sets may emerge relationships that could be explicitly considered in the representation learning process.

We propose a solution for leveraging feature sets into the HGNN architecture, named Feature Set Encoding (FSE). FSE aims at encoding feature set information into the node input vector $\mathbf{x}_v^{\mathcal{FSE}}$. Therefore, we rearrange node input vectors according to the general graph feature superset and concatenate a "bag of features" representation of a node's feature set according to its type. Therefore, we present the first mechanism as FSP, defined superficially as:

$$\mathbf{x}_v^{\mathcal{FSE}} = \mathbf{W} \cdot \text{CAT}\left(\mathbf{x}_v^{\mathcal{C}}, \mathbf{b}_{t_v}\right) \tag{1}$$

where $\mathbf{x}_v^{\mathcal{C}}$ is the rearranged input vector of node $v \in \mathcal{V}$, t_v is the type of node v, $\mathbf{b}_{t_v}$ is the bag of features representation of t_v, and $\mathbf{W}$ is a projection matrix that learns a latent representation of node input vectors to combine both feature values and feature set information. Similarly, we propose two other mechanisms MHAFSE and MHAFSE-F, based on a multi-head attention framework:

$$\mathbf{x}_v^{\mathcal{MHAFSE}} = \frac{\text{softmax}\left(\mathbf{q}_v^{\mathcal{MHAFSE}}, \mathbf{k}_v^{\mathcal{MHAFSE}}\right)}{\sqrt{d}} \mathbf{v}_v^{\mathcal{MHAFSE}} \tag{2}$$

$$\mathbf{q}_v^{\mathcal{MHAFSE}} = \mathbf{W}_q \cdot \mathbf{x}_v^{\mathcal{C}}, \tag{3}$$

$$\mathbf{k}_v^{\mathcal{MHAFSE}} = \mathbf{W}_k \cdot \mathbf{b}_{t_v}, \tag{4}$$

$$\mathbf{v}_v^{\mathcal{MHAFSE}} = \mathbf{W}_{\mathrm{v}} \cdot \mathbf{x}_v^{\mathcal{C}} \tag{5}$$

Vectors $\mathbf{q}^{\mathcal{MHAFSE}}$, $\mathbf{k}^{\mathcal{MHAFSE}}$, and $\mathbf{v}^{\mathcal{MHAFSE}}$ represent the query, key, and value used in the attention mechanism. The query vector seeks relevant information, keys identify where to look, and values hold the actual data. In the multi-head attention approach, the model projects the rearranged input vector as both query and value, while using a bag-of-features representation for the node type as the key, enabling focused attention across multiple feature subspaces. For conciseness's sake, we only mention MHAFSE-F, a third mechanism which uses the concatenation of $\mathbf{x}_v^{\mathcal{C}}$ and $\mathbf{b}_{t_v}$ as input for all query, key, and value vectors.

5. Evaluation & Results

For conciseness's sake, we present in this chapter the outcome of an experiment conducted with the Game Provenance Profile (GPP) dataset on a Player Classification task. We opt for this experiment due to the fact that it shows, as presented in Table 1, that our model PinGLR outperforms both a metrics-based traditional machine learning baseline as well as how our proposed solution FSE enhances the performance of the model in comparison to heterogenous projection matrices architectures found in literature (TSP).

Table 1. Experimental results for the GPP dataset. Values in bold represent the best metric score for the experiment.

Method	Bartle		Dedication	
	Micro F1	Macro F1	Micro F1	Macro F1
Baseline	0.63	0.42	0.65	0.5
TSP	0.69 (1.2e-3)	0.567 (1.02e-2)	0.657 (9.33e-4)	0.547 (1.23e-3)
PinGLR	**0.737 (0.01)**	**0.593 (0.04)**	**0.673 (4.93-e3)**	**0.57 (8.1-e3)**

6. Final Remarks

This work investigates the integration of Game Provenance Graphs (GPGs) and Graph Representation Learning to support Game Analytics and Game Data Science. The main challenges involve addressing game-specific aspects: heterogeneity, dynamicity, and spatiality, with a focus on heterogeneity due to its inadequate handling in prior work. Building on this, the thesis introduces PinGLR, a framework designed to provide end-to-end, task-agnostic learning for Game Analytics, bypassing manual feature engineering. By tackling heterogeneity, we developed Feature Set Encoding (FSE) methods tailored for GPGs. These approaches outperform traditional type-specific projection techniques in both efficiency and representation quality. We also show that PinGLR achieves competitive or superior results compared to traditional metrics-based models.

References

Alencar, V. B. (2020). Prov-dominoes: an exploratory analysis approach for provenance data. Master's thesis, Universidade Federal Fluminense, Niterói.

Angles, R. e Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1–39.

Bengio, Y., Courville, A., e Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Cen, Y., Zou, X., Zhang, J., Yang, H., Zhou, J., e Tang, J. (2019). Representation learning for attributed multiplex heterogeneous network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1358–1368.

Chen, F., Wang, Y.-C., Wang, B., e Kuo, C.-C. J. (2020). Graph representation learning: a survey. *APSIPA Transactions on Signal and Information Processing*, 9:e15.

Drachen, A., Seif El-Nasr, M., e Canossa, A. (2013). Game analytics–the basics. *Game analytics: Maximizing the value of player data*, pages 13–40.

Figueira, F. M., Nascimento, L., da Silva Junior, J., Kohwalter, T., Murta, L., e Clua, E. (2018). Bing: A framework for dynamic game balancing using provenance. In *Proc. of the 17th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, pages 57–5709. IEEE.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., e Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. JMLR.org.

Hamilton, W. L., Ying, R., e Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, 40(3):52–74.

Kazemi, S. M., Goel, R., Jain, K., Kobyzev, I., Sethi, A., Forsyth, P., e Poupart, P. (2020). Representation learning for dynamic graphs: A survey. *Journal of Machine Learning Research*, 21(70):1–73.

Ke, C. H., Deng, H., Xu, C., Li, J., Gu, X., Yadamsuren, B., Klabjan, D., Sifa, R., Drachen, A., e Demediuk, S. (2022). Dota 2 match prediction through deep learning team fight models. In *2022 IEEE Conference on Games (CoG)*, pages 96–103.

Kohwalter, T., Clua, E., e Murta, L. (2012). Provenance in games. *XI SBGames*, pages 162–171.

Kohwalter, T., Oliveira, T., Freire, J., Clua, E., e Murta, L. (2016). Prov viewer: A graph-based visualization tool for interactive exploration of provenance data. In *International Provenance and Annotation Workshop*, pages 71–82. Springer.

Kohwalter, T. C., de Azeredo Figueira, F. M., de Lima Serdeiro, E. A., da Silva Junior, J. R., Murta, L. G. P., e Clua, E. W. G. (2018). Understanding game sessions through provenance. *Entertainment Computing*, 27:110–127.

Kohwalter, T. C., Murta, L. G. P., e Clua, E. W. G. (2017). Capturing game telemetry with provenance. In *2017 16th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, pages 66–75. IEEE.

Le-Khac, P. H., Healy, G., e Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Melo, S. A. (2019). Detecting long-range cause-effect relationships in game provenance graphs with graph-based representation learning. Master's thesis, Universidade Federal Fluminense, Niterói.

Melo, S. A., Kohwalter, T. C., Clua, E., Paes, A., e Murta, L. (2020). Player behavior profiling through provenance graphs and representation learning. pages 62:1–62:11.

Melo, S. A., Paes, A., Clua, E. W. G., Kohwalter, T. C., e Murta, L. G. P. (2019). Detecting long-range cause-effect relationships in game provenance graphs with graph-based representation learning. *Entertainment Computing*, 32:100318.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., et al. (2011). The open provenance model core specification (v1. 1). *Future generation computer systems*, 27(6):743–756.

Myers, S. A., Sharma, A., Gupta, P., e Lin, J. (2014). Information network or social network? the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498.

Skarding, J., Gabrys, B., e Musial, K. (2021). Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 9:79143–79168.

Wang, X., Bo, D., Shi, C., Fan, S., Ye, Y., e Yu, P. S. (2020). A survey on heterogeneous graph embedding: Methods, techniques, applications and sources. *arXiv preprint arXiv:2011.14867*.

Zhang, C., Song, D., Huang, C., Swami, A., e Chawla, N. V. (2019). Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 793–803.