Lego Set Prediction Based on Graphs networks.

1st Jones
Utah State University
Department of Computer Science
Logan, United States
a02248994@aggies.usu.edu

***Abstract*—The increasing popularity of LEGO sets has led to a diverse catalog of intricate designs, each characterized by unique graph structures representing part connectivity and thematic elements. This paper investigates common graph structures within LEGO sets and evaluates methods to predict these structures using artificial intelligence techniques. Three approaches are explored: (1) transforming a global graph of sets into vector embeddings using node2vec, (2) detecting thematic communities via Louvain and spectral clustering, and (3) modeling individual set graphs with metadata extraction followed by random forest classification. The study leverages a comprehensive LEGO dataset to analyze structural patterns and assess the predictive performance of each method. Results highlight the challenges of computational complexity and the efficacy of community detection and classification for thematic prediction, offering insights into scalable graph-based analysis of complex modular designs.**

I. Introduction

The Lego brand has demonstrated remarkable longevity, resulting in an extensive catalog of sets, each comprising intricate 3D structures. This vast collection presents an opportunity to investigate emergent patterns within common themes across these models and to explore whether such patterns can be effectively learned and predicted using artificial intelligence (AI) techniques. This paper proposes and evaluates three distinct methodologies for analyzing and predicting thematic patterns in Lego sets. First, a global graph representation of the sets is constructed and transformed using the node2vec algorithm [1] to capture structural embeddings. Second, community detection is performed using the Louvain method [2] and spectral clustering [3] to identify thematic clusters within the global graph. Finally, an alternative approach is explored, where each Lego set is treated as an individual graph. Metadata is extracted from these graphs, and a random forest classifier [4] is employed to predict the thematic category of each set. The proposed methods are evaluated to assess their efficacy in uncovering and predicting patterns in Lego set designs.

II. Training Data

The dataset utilized in this study is the LEGO Database, provided by Tatman [5]. This comprehensive dataset contains detailed information on official LEGO sets, including parts catalogs, themes, and color inventories for each set. To construct the global graph, set themes were merged with their corresponding parts inventories, forming a unified graph structure. This process enabled the identification of communities of sets based on shared thematic and structural characteristics. For the individual set graph approach, each LEGO set was represented as a fully connected graph, where nodes correspond to individual pieces. Metadata, including the set's thematic category, was extracted from each graph and used to label the graphs for subsequent classification tasks.

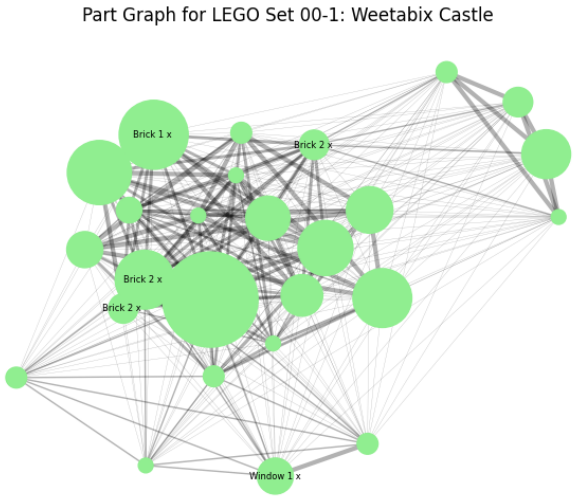

Fig. 1. Example set graph.

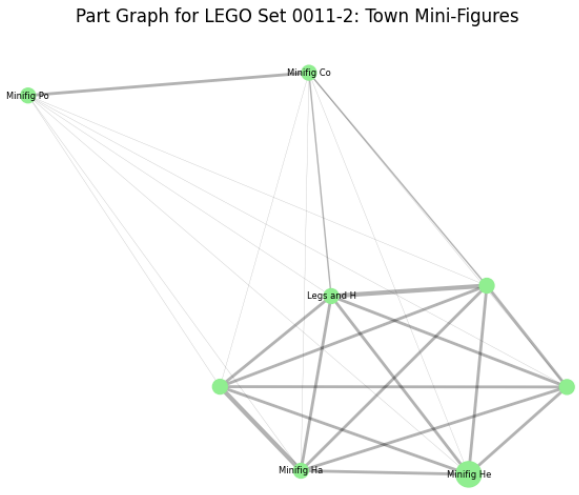

Fig. 2. Example set graph.

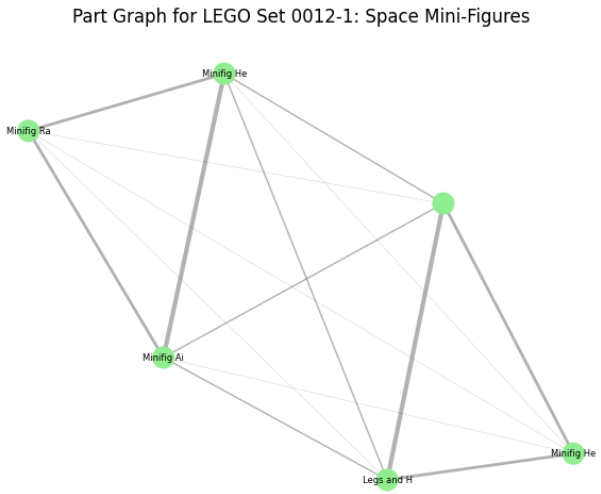

Fig. 3. Example set graph.

The dataset contained a diverse range of LEGO sets, each with varying complexity. For instance, some sets, such as the one depicted in Figure 1, comprised numerous pieces, including many repeated elements. In contrast, other sets, such as those shown in Figure 2 and Figure 3, consisted of only a few pieces.

III. Node2vec-Based Approach

The first approach utilized the node2vec algorithm [1] to transform both the global graph and the collection of individual set graphs into a vector space representation. The objective was to extract community structures from these embeddings to predict thematic patterns in LEGO set constructions. However, this approach encountered significant challenges due to hardware and time constraints. Despite multiple attempts, the algorithm failed to complete execution. Prolonged runs spanning several days resulted in system crashes caused by insufficient memory. Efforts to mitigate these issues, including reducing the number of iterations and downsizing the graphs, were unsuccessful, as the algorithm still did not terminate. Consequently, after evaluating alternative methods, this approach was deemed impractical for the given dataset and computational resources.

IV. Louvain and Spectral Clustering Approaches

This study employed Louvain [2] and Spectral Clustering [3] methods to detect communities within the global graph constructed from the LEGO dataset [5]. These approaches were based on the hypothesis that sets sharing common themes would form distinct communities, while communities common across all sets would be less discriminative. For both methods, multiple models were trained and compared within their respective classes. The best-performing model from each class was then evaluated against a baseline that predicted the most frequent pieces for every set. The evaluation metric assessed the models' ability to predict pieces associated with a given theme. For each theme, a superset of pieces was created by aggregating all pieces from sets of that theme. A predicted piece present in the superset was counted as a true positive (TP). A predicted piece absent from the superset was a false positive (FP). A piece not predicted but present in the superset was a false negative (FN). A piece neither predicted nor present in the superset was a true negative (TN).

Multiple models of each type were evaluated, and the best performers were selected and tested across various themes. The results were averaged over nine trials, as shown in Table I.

TABLE I
Average Performance Metrics for Louvain and Spectral Clustering

Method	Precision	Recall	F1
Louvain	0.017	1.000	0.034
Spectral	0.015	1.000	0.029
Baseline (Most Frequent)	1.000	0.405	0.547

These results reveal key insights. The baseline, which predicts the most frequent pieces, achieved perfect precision (1.000), indicating that a small set of common pieces appears across many sets. However, its moderate recall (0.405) suggests it misses many theme-specific pieces. In contrast, the Louvain and Spectral Clustering methods achieved perfect recall (1.000) but extremely low precision (0.017 and 0.015, respectively), indicating they correctly identified all theme-specific pieces but included many false positives. This resulted in low F1 scores (0.034 and 0.029). The clustering methods successfully detected local communities associated with specific themes but struggled to predict the overall structure of sets due to the limited number of discriminative pieces. Thus, while these methods captured theme-specific patterns, they were less effective at modeling complete set structures.

V. Random Forest Classifier on Individual Set Graphs

This approach modeled each LEGO set as an individual graph using the LEGO dataset [5]. For each graph, metadata was extracted, including the number of nodes, number of edges, average degree, clustering coefficient, density, and average piece quantity. Note that the number of nodes does not directly correspond to the number of pieces due to repeated pieces within a set. This metadata, paired with the set's theme, was used to train a random forest classifier [4] to predict the piece count for each set. After training, supersets were created for each theme by aggregating all pieces present in sets of that theme. To generate synthetic graphs, statistical methods were employed to estimate graph properties (e.g., number of nodes, edges, degree, density, clustering coefficient, and piece quantity) based on the provided theme. The random forest classifier used these properties to predict the piece count, which informed the construction of a random graph. Specific piece assignments were then sampled from the theme's superset to populate the synthetic LEGO set.

This approach successfully generated synthetic LEGO set graphs with desired statistical properties, as illustrated in

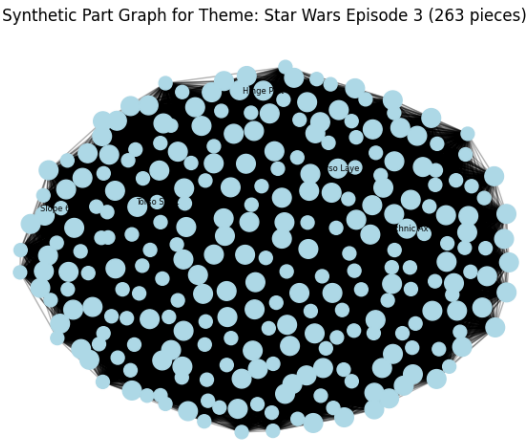

Fig. 4. Example synthetic LEGO set graph.

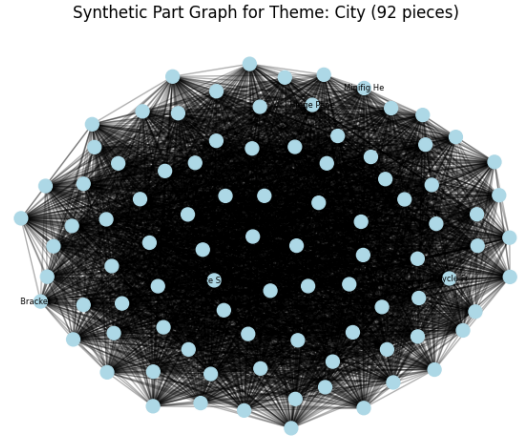

Fig. 5. Second example synthetic LEGO set graph.

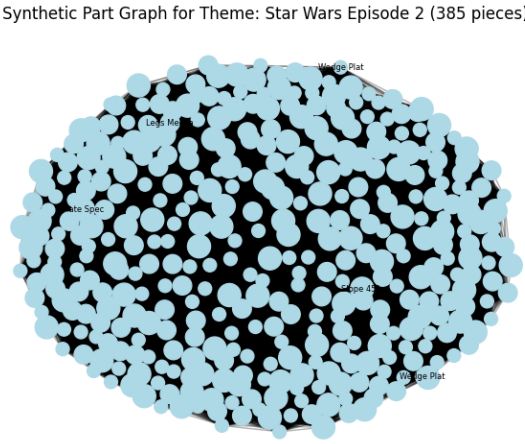

Fig. 6. Third example synthetic LEGO set graph.

Figures 4, 5, and 6. The use of statistical methods ensured that synthetic graphs for the same theme varied, producing unique results per iteration. The random forest classifier effectively learned how piece counts correlate with graph properties across different themes. By sampling from theme supersets, both common and theme-specific pieces were incorporated into the synthetic sets. However, this method did not account for community structures across themes. Consequently, while synthetic sets included relevant pieces, they often lacked contextual coherence. For example, a Star Wars-themed set might include a clone trooper helmet without the accompanying minifigure components, resulting in synthetic sets that lacked logical structure.

VI. Results

This study evaluated three approaches for predicting LEGO set graph structures based on thematic and structural patterns: node2vec (Section III), Louvain and Spectral Clustering (Section IV), and random forest classification on individual set graphs (Section V). The random forest approach was the most effective, as it generated synthetic sets closest to actual LEGO sets in terms of statistical properties.

However, all three methods exhibited limitations, particularly in predicting small-scale sets, such as play packs or minifigure packs, which typically contain 5–10 highly theme-specific pieces. These sets are ideal for community detection due to their lack of common pieces but were consistently overlooked by all methods, which predicted larger sets instead.

The node2vec approach [1] was computationally infeasible, as detailed in Section III, due to excessive memory and time requirements, preventing its practical implementation. The Louvain [2] and Spectral Clustering [3] methods excelled at identifying theme-specific communities, successfully isolating groups of pieces that should remain together. However, as shown in Table I, their low precision (0.017 and 0.015, respectively) indicates they failed to capture the overall context of sets, omitting essential common pieces. The random forest classifier [4] generated statistically accurate synthetic graphs, as illustrated in Figures 4–6, but lacked logical coherence due to its reliance on statistical methods without incorporating cross-theme community structures.

Overall, none of the methods fully achieved the goal of accurately predicting LEGO set structures based on themes and graph communities. While the random forest classifier came closest, its dependence on statistical properties limited its ability to model community-driven contextual relationships, resulting in sets that were structurally plausible but contextually incomplete.

VII. Future Work

The prediction of LEGO set graph structures remains an open challenge with several avenues for improvement. One promising direction is to integrate the random forest classifier (Section V) with community detection methods, such as Louvain [2] or Spectral Clustering [3]. Specifically, the random forest could predict the majority of pieces and graph

properties, while Louvain or Spectral Clustering could enhance community assignments to capture theme-specific piece groupings. This hybrid approach may improve the contextual coherence of synthetic sets, addressing the limitations noted in Section VI.

Additionally, this study modeled LEGO sets as undirected graphs, which limits the representation of assembly relationships. A dataset incorporating directed graph information, such as the order or spatial arrangement of piece connections, could enhance community detection and yield deeper insights into theme-based communities. Such data would enable more accurate modeling of set structures and improve the logical flow of synthetic sets. Future work could also explore advanced graph neural networks to better capture both local and global structural patterns in the LEGO dataset [5].

VIII. Conclusion

This paper investigated three approaches to predict LEGO set graph structures based on thematic and structural patterns using the LEGO dataset [5]: node2vec (Section III), Louvain and Spectral Clustering (Section IV), and random forest classification (Section V). The random forest classifier outperformed the others, generating synthetic sets with statistically accurate graph properties, as shown in Figures 4–6. However, it struggled with contextual coherence due to its reliance on statistical methods without community structures. The Louvain and Spectral Clustering methods excelled at detecting theme-specific communities but failed to capture overall set context, achieving low precision (0.017 and 0.015, respectively, Table I). The node2vec approach was computationally infeasible. All methods struggled to predict small-scale sets, such as minifigure packs, due to dataset imbalance and feature dominance by larger sets.

Despite these limitations, the study provides valuable insights into graph-based analysis of modular designs, highlighting the trade-offs between statistical accuracy and contextual relevance. The proposed methods and their evaluation offer a foundation for future research into hybrid models and directed graph representations, potentially advancing the automated design of complex, theme-driven structures.

References

[1] A. Grover and J. Leskovec, "node2vec: Scalable Feature Learning for Networks," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 855–864, August 2016.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech. Theory Exp., vol. 2008, no. 10, P10008, October 2008.

[3] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Proc. 14th Adv. Neural Inf. Process. Syst. (NIPS), pp. 849–856, December 2001.

[4] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, October 2001.

[5] R. Tatman, "LEGO Database: The LEGO Parts/Sets/Colors and Inventories of Every Official LEGO Set," Kaggle, [Online]. Available: , July 2017. [Accessed: April 21, 2025].