Mining Patterns in BoardGameGeek Data: An Exploratory Analysis for Board Game Recommendation Systems

Eric Oliveira Leal[1], Daniel Villaça de Oliveira Silva[2], Jefferson Elbert Simões[1]

[1]Programa de Pós-Graduação em Informática
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro/RJ – Brasil

[2]Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro, RJ.

{eric.leal}@edu.unirio.br, {daniel.oliveira, jefferson.simoes}@uniriotec.br

***Abstract. Introduction**: The availability of a large-scale dataset from BoardGameGeek (BGG) enables the exploration of board game metadata to extract patterns that can be useful for recommendation systems. However, the high dimensionality and heterogeneity of these attributes complicate the selection of which characteristics have the greatest descriptive power and representativeness. **Objective**: This work aims to analyze this dataset to identify which attributes are most informative and how their relationships can be used to support the design of board game recommendation models. **Methodology**: The approach combined web scrapping techniques to reconstruct the dataset in 2025 with quantitative analyses of Shannon entropy, as well as correlation and conditional expectation analyses on popularity metrics. **Results**: The entropy analysis showed that Categories are the most efficient attribute in terms of exploiting their theoretical potential (86%), whereas Families exhibit higher absolute entropy (9.31 bits) but a highly asymmetric distribution. Mechanics emerge as the central unit of technical description, displaying behavior close to a power law, which is favorable for computationally modeling game profiles. These findings suggest that combining Categories and Mechanics, while carefully handling sparse Families, can provide a solid starting point for defining features and similarity functions in graph-based recommendation systems.*

***Keywords** Board games, BoardGameGeek, Metadata, Data Analysis, Recommendation Systems*

1. Introduction

Board games constitute a complex domain of social interactions and structured mechanics, whose growing popularity has driven the development of scientific studies on categorization methods and data-driven recommendation systems [Xiao 2025, Kim et al. 2020]. However, given the vast range of metadata available to describe a game, from time and age metrics to complex mechanics and themes, a fundamental challenge emerges in the analysis of board games which is identifying which attributes possess the greatest descriptive power and representativeness for robustly characterizing a title.

The difficulty in defining a board game through its data lies in the dimensionality and potential redundancy of its attributes. It is often assumed that category is the primary descriptor, but practical experience shows that attributes such as perceived complexity or the combination of specific mechanics may carry significantly greater descriptive weight in defining a game's identity [Xiao 2025]. For instance, within the card game category, traditional games such as Bridge and more modern ones such as Uno, despite sharing the same primary category, exhibit substantial differences in mechanics, complexity, and interaction dynamics, which can lead to entirely distinct user experiences.

Recent studies indicate that, for the development of efficient computational models, it is imperative to distinguish which variables constitute noise and which are, in fact, the determining characteristics that define a game's profile within the digital ecosystem [Xiao 2025, Putra e Wibowo 2022]. Modern recommender system techniques have increasingly evolved beyond traditional collaborative filtering, incorporating advanced approaches such as deep learning-based embeddings, graph neural networks, and hybrid content-aware architectures to capture complex domain dynamics. However, the performance and interpretability of these sophisticated models remain heavily dependent on the quality and representativeness of the underlying feature space.

In this context, the present work aims to analyze a large-scale dataset from BoardGameGeek, an online social network dedicated to board games, in which users interact through reviews, ratings and preferences. An up-to-date and large-scale dataset was built by using web scraping techniques, which this work uses for intending to identify patterns and relevant relationships among attributes that may yield meaningful and applicable insights for modern recommendation systems. Through statistical and exploratory analyses, the goal is to understand which characteristics are most informative, how they correlate, and how they can be leveraged to improve the representation and recommendation of board games.

2. Related Works

Understanding a board game dataset depends on clearly defining its constitutive attributes, how they are organized into taxonomies, and how they are used by computational methods. In the context of BoardGamesGeek (BGG), this issue is particularly relevant because metadata not only describe a game but encapsulate cataloging conventions, design choices and community usage practices. Therefore, before employing such attributes in similarity or recommendation tasks, it is necessary to examine the meaning of each variable, its relationship with game design and the limitations imposed by the structure of the dataset itself [Piette et al. 2021, Samarasinghe et al. 2021].

The game design literature shows that board games can be analyzed through sets of general concepts and specific mechanisms. Silva et al. (2025) discuss the consolidation of design methods applied to the board game domain, emphasizing the importance of conceptual frameworks to organize elements of creation and analysis. In addition, Piette et al. (2021) propose a more systematic vocabulary to represent general game properties, including structural, temporal, informational and rule-based aspects, which is particularly useful for formalizing attributes in empirical datasets.

This concern with game representation also appears in studies focused on mechanics, which should not be treated as isolated components, as they interact with each other and relate to dimensions such as complexity and game structure [Samarasinghe et al. 2021]. Moreover, mechanics can be classified and interpreted as relevant analytical units, contributing to understanding how games are structured and how their attributes can be encoded in databases such as BGG [Lima e Falcão 2025].

Harrington (2024) argues that communities on BGG not only consume games but also perform identities, produce knowledge and participate in processes of platformization an replatformization that shape circulation of games and their meanings. This is particularly relevant for BGG because its attributes, ratings and classifications reflect not only intrinsic properties of games but also social practices and mediation patterns constructed by users and online communities.

From a computational perspective, these works indicate that the analysis of board games requires a careful reading of data structures. Game recommendation in digital platforms depends on the quality of attribute representation and the ability to capture relationships among games, users and usage contexts. Kim et al. (2020) show that sequential data from board game platforms can be used for personalized recommendations with deep learning techniques, reinforcing the importance of well-structured and semantically consistent datasets. At the same time, context plays a crucial role in game recommendation, indicating that game attributes and situational information can be combined to improve the retrieval of items suited to user profiles [Jorro-Aragoneses et al. 2024].

In this scenario, analyzing BGG goes beyond simple metadata extraction, highlighting that the dataset should be understood as a representation system in which attributes such as mechanics, categories, duration, number of players and complexity operate in an articulated manner to describe games and support subsequent analytical applications. The literature suggests that some these fields are more informative or stable than others, requiring methodological care in selecting the variables to be explored [Samarasinghe et al. 2021, Piette et al. 2021]. Thus, exploratory data analysis should simultaneously consider the formal structure of attributes, the social logic behind data production and the role of mechanics as central units for description and comparison [Lima e Falcão 2025, Harrington 2024].

Although the social nature of BGG is acknowledged, the literature still lacks a systematic analysis of the descriptive power, redundancy, and interaction effects of different BGG attributes, especially when considering large-scale datasets. While mechanics and categories are widely used, their relative importance and structural interactions are not consistently evaluated across the literature. This work aims to bridge this specific gap by providing a comprehensive exploratory assessment.

3. The BGG database

BoardGameGeek (also known as BGG) is an online forum for board game players which launched in 2000 and has since amassed over two million registered users. It is self-described as "an online resource and community that aims to be the definitive source for board game and card game content"[1]. While lacking features present in other board game-related platforms such as an online game environment (such as Tabletopia or Board Game

[1] https://boardgamegeek.com/wiki/page/Welcome_to_BoardGameGeek

Arena) or an established online marketplace[2] (such as Ludopedia), its main differential is its crowdsourced game database, which currently comprises over 300,000 games. This strategy was fundamental in establishing BGG as the biggest platform of the genre since its launch, by providing useful information to their userbase while also fostering a sense of community via its forum-related features.

BGG provides two versions of an XML API that allow app developers to retrieve select data from its database in order to build additional applications, although their Terms of Service restrict its usage to non-commercial applications and explicitly excludes training of AI models, including LLMs. Their usage is subject to rate limits controlled by a standard *application token* scheme.

Our first main goal is to build a crawler that will harvest publicly available data from the BGG database, thus allowing us to create a board game dataset[3] and statistically analyze it. For the remainder of this section, we describe the publicly-facing data model of the BGG database and our web scraping process.

BGG provides limited access to its data through a small number of endpoints. The fundamental concept is that of a *thing*, which can represent any physical product, ranging from board games and their expansions to accessories and magazines. Each object has a unique identifier (ID), and the main API endpoint for our purposes allow for retrieval of several attributes of one of more *things* by using their IDs as key. The *type* attribute classifies each *thing* into one of a few classes, including board games and board game expansions. While this endpoint permits filtering by one or more of these types in a query by ID range, *things* of all types share a global ID range. Thus, scraping for all board games in the database and their respective attributes requires a large series of queries to this endpoint, as the API bounds applications to at most 20 items per request and employing a filter only decreases the response size.

Each *thing* is associated with additional information, retrieved as inner XML tags; these include primary and alternate names (e.g. in foreign languages), description, year of publishing, cosmetic data such as a reference image and a thumbnail, and crowdsourced recommendations for number of players and playing time. While we preclude the full list of available data for brevity, for the purposes of our analysis, the following properties of a game are particularly relevant:

Categories: user-defined "means with which to group games based on like subjects or similar characteristics", including[4] activity Categories such as `Civilization` and `Educational`, thematic Categories such as `Medieval` and `Humor`, and skill Categories such as `Trivia` and `Bluffing`.

Mechanics: a specific functional aspect of a game's rules. Notable examples[5] include `Auction/Bidding`, `Card Drafting`, `Grid Movement`, `Modular Board`, and `Worker Placement`.

[2]BGG launched a marketplace called GeekMarket in 2022, though it presents itself as a distinct Web environment amid links to other marketplaces such as eBay and Amazon. Thus, we treat BGG and GeekMarket as separate platforms with a joint owner.

[3]Due to the BGG TOS restrictions, despite all scraped data being publicly available, we do not currently have the license to publicly share the consolidated dataset.

[4]See full list at `https://boardgamegeek.com/wiki/page/Category`.

[5]See full list at `https://boardgamegeek.com/wiki/page/mechanism`.

Families: specific groupings of 3 or more games which are "submitted by users [and] share something from theme to historical period, from a character to a sport" or, for instance, comprise especially designed series or collections. Examples include `Carcassonne` (comprising the Euro-style game Carcassonne, its expansions, reimplementations, and fan variations, for a total of 188 items) and `Game: Werewolf / Mafia` (comprising the game Werewolf and other which follow its general rule structure, such as The Resistance, for a total of 360 items).

We note that each game is typically associated to multiple Categories, Mechanics, and Families. Therefore, in our analysis (see Section 4), we will refer to *assignments* of games to each of these properties, as we will want to treat each of them as a random variable and analyze their statistical properties.

Other relevant properties available through this endpoint are statistics relating each game to its BGG userbase, including:

Rating statistics: users can rate each game in an integer, 1–10 scale, with both amount of users with an assigned rating and the aggregated, average rating (measured to one decimal place) available;

Listing statistics: users can add games to several personally curated lists, including: *owned* (games one owns), *wanting* (games one wants to buy or own), and *wishing* (games one wishes to play). The aggregate number of users who added a game to each of these lists is available separately for each game.

Altogether, the amounts of users who rated a game (`NumUserRatings`), own a game (`NumOwned`), want to buy/own a game (`NumWant`), and wish to play a game (`NumWish`) can be construed as different measures of popularity, reflecting subjective factors such as design quality and marketing campaigns.

Our crawling process occurs in two stages. First, we sent successive requests to this endpoint up to the highest ID (which is given to the most recently registered *thing*) and storing their respective raw responses in multiple XML files. Then we reprocess this raw data to discard irrelevant entities (i.e., those whose type is not "board game" or "board game expansion"), and remove fields for each game that provide information outside of the scope of this research, while reserializing the resulting data in a single JSON file for easier future manipulation.

The resulting dataset is a snapshot of all board games and expansions registered in BGG, as collected by November 2025. Overall, we collected information for a total of 169,633 board games and expansions.

4. Analysis

In order to understand how games are distributed throughout existing Categories, Mechanics, and Families, we study the distribution of games by each of these properties separately. We also quantify the informational content of each of these properties by calculating their *Shannon entropy* in bits, as given by Equation 1 [Shannon 1948]:

$$H(X) = -\sum_{x} p_x \lg p_x, \tag{1}$$

where X is an attribute denoted as a random variable, x spans all possible values of X, p_x is the relative proportion of game assignments to attribute x, and $\lg$ denotes the binary logarithm function.

We are also interested in analyzing how often two distinct values of these attributes are assigned to the same game, which we call a *co-occurrence*. This can be calculated by building, for any value v of a Category, Mechanic, or Family, a bit array A_v indicating the assignment of each game to v. Given two values v, w, the *number of co-occurrences* of v and w is given by the dot-product $A_v \cdot A_w$, and the *proportion of co-occurrences* is given by $\|A_v \wedge A_w\|/\|(A_v \vee A_w)\|_1$, where $\wedge$ and $\vee$ denote, respectively, bitwise AND and OR operations. This can be contrasted with the *expected proportion of co-occurrences* of both values, which given an assumption of independence, can be calculated as $\|A_v\|_1 \cdot \|A_w\|_1/n$, where n is the total number of games. This analysis is not particularly relevant for Families as they tend to have more specific meanings, thus we restrict it to Categories and Families.

Regarding the popularity metrics, it is interesting to understand how they relate to a game's aggregated rating (as a representation of users' opinion about that game) and to the belonging of games to specific Categories, Mechanics, and Families. This is achieved by analyzing the distribution of games' aggregate ratings conditioned on their popularity.

5. Results

We begin by analysing the distributions of assignments for Categories, Mechanics, and Families. Figure 1 presents all three distributions in CCDF form, with some essential statistics in Table 1 and calculated entropy for each distribution in Table 2.

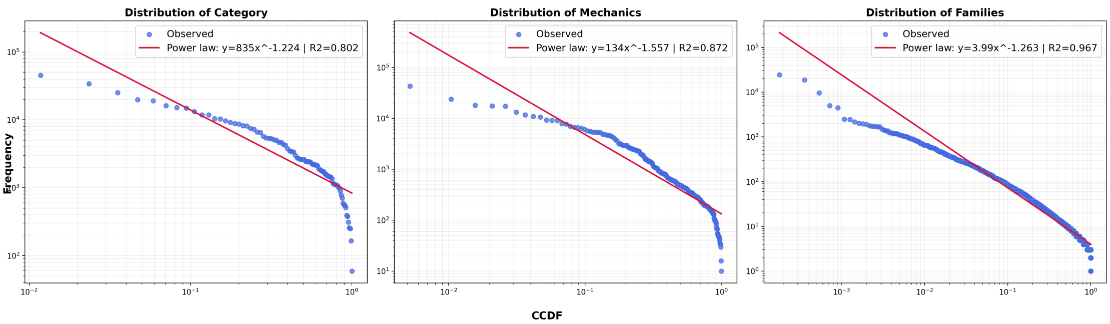


Figure 1. Complementary cumulative distribution of the characteristics

Table 1. Assignments for each attribute

Attribute	Values	Assignments	Avg. assignments/value	Avg. assignments/game
Categories	85	463,529	5,453.28	2.73
Mechanics	192	441,270	2,298.28	2.60
Families	5533	346,690	62.66	2.04

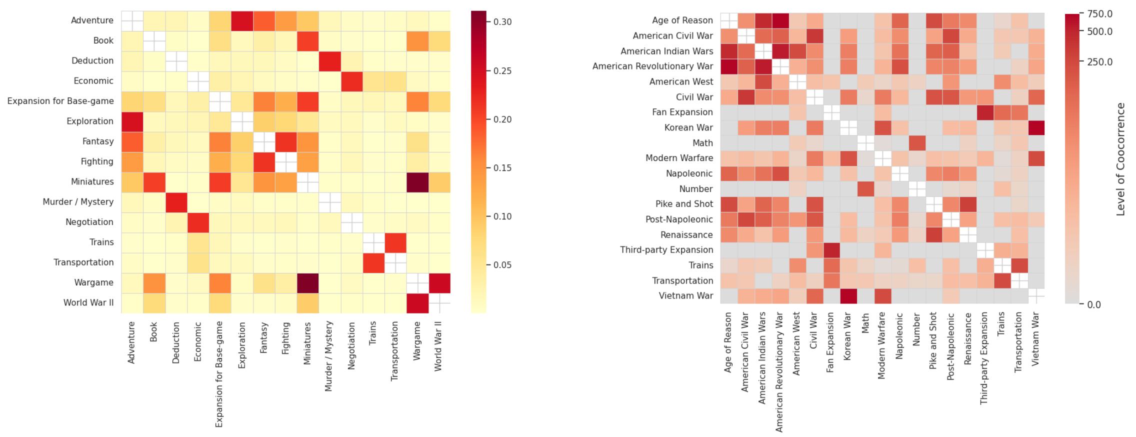
As can be seen in Figure 1, all three attributes exhibit features of long-tailed distributions, with values spanning multiple orders of magnitude with non-trivial probability. The distribution of assignments for Families, in particular, strongly resembles a power-law distribution considering its Determination Coefficient (R^2) equals 0.967.

Table 2. Entropy results for each attribute

Attribute	Values (N)	Entropy H (in bits)	$\log_2(N)$ maximum	$H/\log_2(N)$
Categories	85	5.5504	6.41	86.6%
Mechanics	192	6.0954	7.58	80.4%
Families	5533	9.3166	12.43	74.9%

This heavy inequality in all distributions is also displayed in their high relative entropies, as indicated in Table 2. These results, however reveal the coexistence of two behavioral patterns. From an absolute value perspective, the Families attribute exhibits the highest entropy, indicating a richer and more numerically diverse information space than the other attributes. Meanwhile, when considering the theoretical maximum, the order is reversed, as Categories reach over 80% of the maximum entropy achievable for their vocabulary, compared to the other attributes. Thus, despite having the largest vocabulary, the Families attribute presents a more concentrated distribution relative to its potential, and its high entropy seems to be mostly originating from the large amount of available Families.

Figures 2a and 2b show the co-occurrence rates for each pair of Categories, both relative to each other and to the expected co-occurrences under independence assumptions. It is notable that this matrix is extremely sparse, with most Category pairs exhibiting zero or negligible co-occurrence. This finding reinforces the long-tail pattern, as only a few dominant Categories co-occur frequently, while many others remain largely isolated.



(a) Co-occurrence matrix of Categories, normalized by number of games. Categories with no co-occurrence values $\geq$ 0.15 were omitted for simplicity (14 out of 85 displayed).

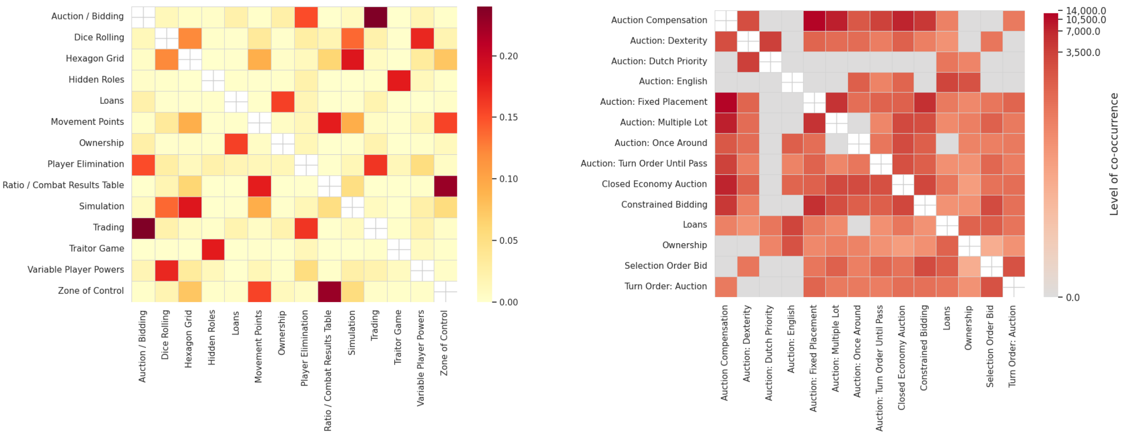
(b) Matrix of ratios between observed and expected co-occurrence for Categories. Categories with no ratios $\geq$ 150 were omitted for simplicity (19 out of 85 displayed).

Figure 2. Excerpts of co-occurrence matrices for Categories.

The analysis of ratios between observed and expected co-occurrence reveals high heterogeneity of the data. Pairs with strong positive associations reach proportions hundreds of times higher than would be expected by chance, with particularly strong examples found in domains of historical conflict such as `Napoleonic`, `World War II` and `Wargame`, or in editorial contexts such as `Expansion for Base-game` and `Third-party Expansion`. At the opposite end, structural incompatibilities between genres appear, as Categories such as `Children's Game` and `Wargame`, or `Educational` and `Horror`, rarely share the same title, reflecting incompatible design constraints and mutually exclusive target audiences.

The visualization in Figure 2b shows that a large amount of pairs exhibit some degree of association above independence. This suggests a structure composed of two overlaid layers: a strong, localized one, restricted to a few semantic clusters, and a weak, diffuse one that permeates the entire Category space. This behavior possibly reflect biases introduced by the platform's collaborative curation practices. Interestingly, there is relatively little intersection between those Categories showing high relative co-occurrence values and those showing high absolute co-occurrence values, the most notable exceptions being the `Trains` and `Transportation`, which naturally have high co-occurrence values with each other by both standards. This indicates that high co-occurrences by frequent Categories are mostly due to these high frequencies and random fluctuations, while most rarely used Categories are more likely to fulfill niche markets in combination with one another.

Meanwhile, Figures 3a and 3b show these same plots for the Mechanics attribute. Again, we see a highly sparse matrix indicating that most Mechanic pairs occur rarely, while a limited number of combinations concentrate most of the overlap. Some noteworthy pairs highlight specific patterns as board games follow given gameplay logics, e.g., `Auction/Bidding` tend to be highly related to `Trading` ones. Likewise, Mechanics such as `Loans` and `Ownership` frequently appear in financially oriented games, while `Hexagon Grid`, `Simulation`, `Dice Rolling`, `Movement Points` and `Zone of Control` form a separate cluster, usually associated to Wargames. In contrast, several mechanics remain isolated, reflecting they fill a more specialized use.



(a) Co-occurrence matrix of Mechanics, normalized by number of games. Mechanics with no co-occurrence values $\geq$ 0.15 were omitted for simplicity (14 out of 192 displayed).

(b) Matrix of ratios between observed and expected co-occurrence for Mechanics. Mechanics with no ratios $\geq$ 1500 were omitted for simplicity (14 out of 192 displayed).

Figure 3. Excerpts of co-occurrence matrices for Mechanics.

It is clear when we analyze Figure 3 that some Mechanic pairs reach very high observed-to-expected ratio, that means certain Mechanics are structurally dependent on

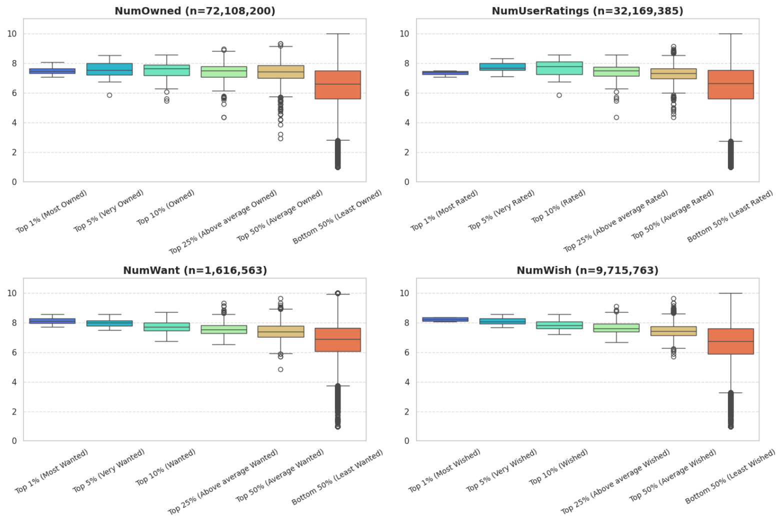


Figure 4. Boxplots for average rating by popularity bucket

one another within particular game archetypes. `Auction` variants, for example, exhibit a strong mutual reinforcement. At the same time, many cross-domain pairs remain near the baseline expectation, suggesting that the design space of Mechanics is constrained by thematic or functional compatibility.

Regarding popularity, Figure 4 shows the distribution of aggregate ratings for several popularity percentiles. This figure highlights how tiers of high popularity exhibit a consensus-driven distribution, with narrow interquartile ranges and showing low agreement among the users and higher median values except for the top rungs in two of the popularity metrics. Furthermore, as popularity decreases, the distributions of ratings becomes more dispersed. In particular, the bottom half games in popularity, by all metrics, display an extremely high proportion of outliers by aggregate rating, both towards higher and (especially) lower rating values, suggesting both acclaimed titles and a poorly received mass of games among those lesser-known.

Complementarily, Figure 5 presents the expected aggregate rating of games, conditioned on their popularity exceeding a given threshold. Our analysis indicates that games owned by more people and with a higher number of ratings tend to have higher average scores, suggesting that exposure and community engagement are directly associated with users' perception of quality.

6. Conclusion and future work

This work conducted a quantitative investigation of the BoardGameGeek dataset, aiming to derive conclusions that can ground and guide the development of board game recommendation systems. The entropy showed that, although the Families attribute has the largest information space (9.31 bits), its distribution is highly asymmetric and long-tailed, concentrated in niche terms. In contrast, the Categories attribute proved the most efficient in exploiting its theoretical potential (86%), suggesting a more stable and consolidated vocabulary for the general description of titles.

The co-occurrence analyses revealed a sparse data structure, yet with the presence of strong semantic clusters such as historical war games and expansions, along with

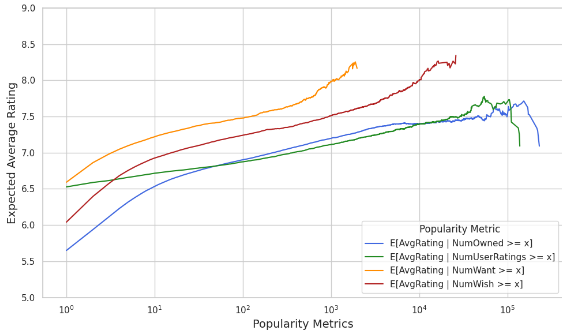


Figure 5. Expectation of average rating, conditioned on values of popularity metrics.

design-driven incompatibilities, such as the repulsion between children's games and complex strategy games (`Wargames`). From the perspective of popularity, it was observed that community engagement and game exposure are positively correlated with perceived quality, evidenced by the tendency toward higher average ratings in games with larger numbers of owners and ratings.

These results show that Mechanics act as the central unit of technical description, approximating a power-law-like pattern that facilitates the computational modeling of game profiles. Moreover, the choice of attributes for machine learning models should consider the redundancy and sparsity of Families, prioritizing Categories and Mechanics to ensure a better generalization. Building on these findings, the attribute space can be organized more effectively for a recommendation system by using Categories and Mechanics as the central axes of a characterization, treating Families as a niche signal and exploiting co-occurrence patterns to define semantic groups and avoid incoherent recommendations.

A natural next step is to explore more deeply the relationships among users and between users and games. However, crawling the data for this task is considerably more challenging as the BGG API only provides endpoints for user data using their usernames as key, and BGG does not provide an explicit user list[6]. A possible strategy in this case is to extract usernames directly from game information such as ratings and reviews, then use them as seeds in a process of snowball sampling through the user friendship network. While this might not reach all existing users, it might be enough to provide a large enough fraction of this network.

In future work, we intend to employ these conclusions to design recommendation algorithms for board games, exploiting playing patterns and game similarities in the process. A full analysis of this dataset, including user data and user-game relationships, will be crucial in this development. Additionally, future work could investigate whether the observed sparsity patterns remain stable when restricting the analysis exclusively to highly rated or highly owned games, as well as explore the relationship between the number of awards a game has won and other attributes present in the database.

[6] We got in touch with BGG and requested it. They refused.

References

Harrington, J. (2024). (re)platforming subcultures: Sinosphere board game geeks. *Global Media and China*, 0(0):20594364241310553.

Jorro-Aragoneses, J. L., Cantador, I., e Bellogín, A. (2024). Context-aware board game recommendations. *Bari, Italy), RecSys*.

Kim, J., Wi, J., Jang, S., e Kim, Y. (2020). Sequential recommendations on board-game platforms. *Symmetry*, 12(2).

Lima, H. e Falcão, T. (2025). Relações entre mecânicas de jogos de tabuleiro e os pilares do pensamento computacional. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, pages 1570–1579, Porto Alegre, RS, Brasil. SBC.

Piette, É., Stephenson, M., Soemers, D. J. N. J., e Browne, C. (2021). General board game concepts. *CoRR*, abs/2107.01078.

Putra, D. e Wibowo, A. (2022). Sentiment analysis for board game review using deep learning and sentiment lexicon. *Int. J. Emerg. Technol. Adv. Eng*, 12(6):56–62.

Samarasinghe, D., Barlow, M., Lakshika, E., Lynar, T., Moustafa, N., Townsend, T., e Turnbull, B. (2021). A data driven review of board game design and interactions of their mechanics. *IEEE Access*, 9:114051–114069.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Silva, E., Araújo, C., e Júnior, J. S. (2025). Métodos de game design aplicados a jogos de tabuleiro: Uma revisão sistemática. In *Anais do XXIV Simpósio Brasileiro de Jogos e Entretenimento Digital*, pages 219–232, Porto Alegre, RS, Brasil. SBC.

Xiao, T. (2025). Analysis of factors influencing board game ownership based on a gradient boosting model. *Theoretical and Natural Science*, 130:16–24.