A Bayesian Perspective on Modelling Sparse User Preference Structures

Tony Livins

Solent University

United Kingdom

Contents

Abstract

Contemporary recommender systems depend heavily on user preference data to identify, interpret, and predict items of interest. Despite its importance, acquiring and disseminating such data presents significant challenges across various domains. Data collection can be resource-intensive, particularly when large and diverse user populations are required, and concerns surrounding privacy often restrict the sharing of authentic user information. To address these limitations, this study introduces a novel probabilistic framework for the generation of realistic synthetic preference data.

The proposed model constructs an initial dense representation of user–item interactions through matrix factorization, which is subsequently transformed into a sparse structure to better reflect real-world conditions. Crucially, the approach integrates key aggregate characteristics of the source data, including patterns of user engagement and item popularity, as well as the dependencies between them. This enables the generation of synthetic datasets that preserve the statistical properties observed in genuine user behaviour.

Empirical evaluation demonstrates that the model is capable of closely approximating real-world datasets across multiple domains, achieving high fidelity according to a range of quantitative metrics. The framework offers practical value for both research and applied settings, supporting the creation of new datasets and the augmentation of existing ones. In doing so, it facilitates robust model evaluation and provides an effective strategy for overcoming data scarcity through enhanced bootstrapping techniques.

1 Introduction

User preference data has emerged as a critical asset in contemporary decision-making, underpinning applications across industry, government, and digital platforms. Such data enables systems to model behaviour, anticipate user needs, and deliver personalised recommendations. Despite its strategic importance, obtaining high-quality preference data remains a complex and resource-intensive process. Effective data collection typically requires access to a sufficiently large user base, a well-defined catalogue of items, and reliable mechanisms for capturing user interactions, such as ratings or click behaviour. In many cases, these requirements introduce substantial financial, technical, and logistical constraints.

Beyond the challenges of acquisition, the dissemination of user preference data is frequently restricted. Privacy concerns, ethical considerations, and regulatory frameworks often limit the extent to which real-world datasets can be shared or reused. As a result, research and development in recommender systems tend to rely on a small number of publicly available benchmark datasets. While these datasets have played a significant role in advancing the field, an over-reliance on them can constrain methodological innovation and reduce the generalisability of findings.

In other areas of artificial intelligence, synthetic data generation has proven to be an effective strategy for overcoming similar limitations. Fields such as planning and reinforcement learning have long benefited from established repositories of simulated environments and benchmark datasets, enabling rapid experimentation and comparative evaluation. Likewise, synthetic data has become a standard tool in domains including social network analysis and statistical modelling, where it facilitates the study of complex systems under controlled conditions. By extending this paradigm to recommender systems, synthetic preference data offers the potential to broaden experimental scope, support reproducibility, and enable research in settings where real data cannot be accessed due to privacy or ethical constraints.

This work introduces a novel probabilistic framework for generating synthetic user preference data. The proposed approach builds upon probabilistic matrix factorization, a well-established technique for modelling latent user–item relationships. However, rather than using this model solely for prediction, it is adapted to support data generation.

Specifically, an initial dense user–item interaction matrix is produced and subsequently transformed through a structured masking process. This masking mechanism governs which interactions are retained, and is parameterised by factors such as user activity levels, item popularity, and their interaction effects. The framework can also incorporate auxiliary information, allowing it to capture more complex behavioural dependencies.

Although the model is flexible and capable of producing data under a range of configurations, its effectiveness is demonstrated through empirical evaluation on datasets spanning multiple domains, including film, literature, music, and e-commerce. The generated data is shown to replicate key statistical properties of real-world datasets with high fidelity. Furthermore, the synthetic datasets can be directly utilised by existing recommendation algorithms, demonstrating their practical applicability. This enables not only the creation of entirely new datasets but also the extension of limited data resources, supporting more robust experimentation and model development.

2 Related Work

The generation of synthetic user preference data has received comparatively limited attention within the recommender systems literature, despite its clear relevance to both research and practical applications. Early conceptual work by Pasinato et al. outlines a potential architecture for simulating preference data through the explicit construction of user and item profiles. In this framework, distributions governing the number of items evaluated per user, as well as the corresponding ratings, are defined directly. A key advantage of this approach lies in its capacity to prescribe the statistical properties of ratings through predefined probability density functions. However, the absence of an implemented system limits its empirical validation and practical utility.

In contrast, the framework proposed in this study adopts a similar high-level objective but differs fundamentally in its modelling strategy. Rather than specifying observable distributions directly, the approach operates through latent representations of users and items. This enables a more flexible and scalable mechanism for capturing complex interaction patterns, allowing the structure of the generated data to emerge from underlying probabilistic relationships rather than being imposed explicitly.

An alternative and more established approach to synthetic data generation is based on clustering techniques. Tso and Schmidt-Thieme introduce a method in which both users and items are partitioned into latent groups, forming user clusters and item clusters respectively. Each user and item is probabilistically assigned to a cluster through Dirichlet-distributed priors, reflecting the variability in user behaviour and item characteristics. The interaction structure between these clusters is then modelled through a conditional distribution that captures the likelihood of associations between user and item groups.

To regulate the complexity and informativeness of the generated data, this method employs an entropy-based constraint. Specifically, the conditional entropy of user clusters given item clusters is iteratively adjusted through repeated sampling from a modified chi-squared distribution, with values exceeding a defined threshold discarded. This process continues until a target entropy level is achieved, thereby controlling the degree of uncertainty and diversity in the resulting dataset.

Among existing methods, this clustering-based approach represents one of the few that has been both implemented and empirically evaluated for synthetic preference data generation. As such, it provides a suitable benchmark for comparison. In this work, it is adopted as a baseline in order to assess the effectiveness of the proposed probabilistic framework in capturing the statistical properties of real-world preference data.

$$H(C^U|C^I) = -\sum_{i=0}^{|C^U|}\sum_{k=0}^{|C^I|} \frac{P(C^I = k, C^U = i)\log_2(P(C^U = i|C^I = k))}{\log_2(|C^I|)}.$$

Following the formation of cluster-level relationships, users generate item interactions by sampling from item clusters according to the learned conditional distribution. Specifically, the likelihood that a user i, belonging to user cluster C, selects items from an item cluster C is governed by the conditional probability. Once an item cluster has been selected, individual items within that cluster are sampled according to a binomial process, where the

associated parameter controls the probability of selecting each item. This procedure results in the construction of a binary user–item interaction matrix, denoted R, representing observed preferences.

Although this clustering-based framework is flexible and broadly applicable, it suffers from limited interpretability. In particular, the relationship between model parameters and the statistical properties of the generated data is not straightforward, as the underlying probability distributions are not explicitly defined. In contrast, the approach proposed in this work preserves interpretability by directly specifying the probability density function while simultaneously enabling the generation of structured attribute information. This is achieved by modelling preference data as a function of latent item attributes, allowing for more transparent control over the resulting data characteristics.

An alternative perspective frames the generation of user preference data as a random graph construction problem. Under this view, the data can be represented as a bipartite graph, where users and items form two disjoint sets of nodes, and edges correspond to observed interactions. A wide range of random graph models have been developed, with early contributions such as the Erdős–Rényi model providing foundational insights. However, such models are overly simplistic for capturing the complexities of user preference behaviour. More advanced approaches allow for the specification of structural properties, including degree distributions and sparsity patterns.

For example, Caron and Fox propose a framework for generating sparse and exchangeable bipartite graphs, though it does not offer direct control over the degree distribution. Similarly, Newman et al. introduce a method for constructing bipartite graphs with arbitrary degree distributions, but their approach does not incorporate latent user–item interactions or account for budget constraints, both of which are central to realistic preference modelling. While these graph-based methods have not been explicitly applied to recommender system data generation, they provide a useful theoretical foundation that informs the design and interpretation of the proposed model.

3 A Sparse Probabilistic User Preference Model

The objective of the proposed framework is to generate a user–item preference matrix $R \in$

$\mathbb{R}^{N\times M}$, suitable for use in recommender system applications. Each row of Rcorresponds to an individual user, while each column represents an item. Entries within the matrix encode user preferences, which may be derived either implicitly through behavioural signals such as purchases, views, or interactions, or explicitly through rating mechanisms. Consistent with standard practice in recommender systems, a value of zero indicates the absence of an observed interaction between a user and an item. In real-world scenarios, such matrices are typically highly sparse, reflecting the limited number of interactions relative to the total number of possible user–item pairs. Nevertheless, the proposed framework is sufficiently flexible to generate datasets with varying levels of sparsity, including fully dense representations if required.

The Sparse Probabilistic User Preference (SPUP) model generates synthetic preference data through a two-stage process. In the first stage, a dense latent preference matrix is constructed using a probabilistic matrix factorization approach. This matrix captures the underlying affinity between users and items, representing the degree to which a user is expected to favour a given item. In the second stage, this dense representation is transformed into a sparse structure by introducing user-specific constraints. These constraints are operationalised through the concept of user budgets, which determine the number of items a user is likely to interact with. Items are then sampled from a probability distribution conditioned on both the user's latent preferences and global item popularity, resulting in a realistic and structured sparsification process.

Probabilistic Matrix Factorization

The proposed model builds upon probabilistic matrix factorization (PMF), a widely adopted generative framework in recommender system research. PMF models the observed preference matrix as the product of two lower-dimensional latent factor matrices, which are typically interpreted as user preference vectors and item attribute vectors. In conventional applications, these latent factors are learned by minimising the reconstruction error over observed entries in the data.

In contrast to standard PMF implementations, where latent variables are inferred from existing datasets, the present approach samples these latent representations directly from predefined distributions. This enables the generation of synthetic data without reliance on observed interactions. The resulting dense preference matrix, denoted $\tilde{R}$, is therefore modelled as a noisy product of latent user preferences and item attributes, capturing the inherent variability and uncertainty present in real-world user behaviour.

$$\mathbf{U}_i \sim \mathcal{N}(\mathbf{0}, \sigma_u^2\mathbf{I}) \quad \text{and} \quad \mathbf{V}_j \sim \mathcal{N}(\mathbf{0}, \sigma_v^2\mathbf{I}) \tag{1}$$

$$\widetilde{\mathbf{R}}_{ij} \sim \mathcal{N}(\mathbf{U}_i^\top \mathbf{V}_j, \sigma_p^2), \tag{2}$$

In this formulation, U denotes the latent preference vector associated with user i, while V represents the latent attribute vector corresponding to item j. The identity matrix is denoted by I, and the parameter σ controls the variance of the Gaussian noise component, capturing uncertainty in the interaction process.

Generation of the Mask Matrix

The Aldous–Hoover representation theorem establishes that matrices generated under exchangeable probabilistic models, such as probabilistic matrix factorization, tend to be either fully dense or trivially empty. As a consequence, the initial preference matrix, obtained from the PMF stage, typically exhibits a dense structure. However, this characteristic does not align with real-world user–item interaction data, which is inherently sparse due to the limited number of interactions each user has with the overall item set.

To address this discrepancy, the second stage of the proposed framework introduces a sparsification mechanism designed to transform the dense matrix into a more realistic representation. This is achieved through the construction of a mask matrix that selectively retains a subset of interactions. The sparsification process consists of three key steps. First, a user-specific budget is sampled to reflect the activity level of each individual, thereby determining the number of interactions they are expected to generate. Second, a popularity score is assigned to each item, capturing the global tendency of items to be selected. Third, for each user, items are sampled according to a probability distribution that incorporates both the user's latent preferences, as derived from the PMF stage, and the overall popularity of the items, subject to the constraint imposed by the user's budget.

To ensure that the resulting data exhibits realistic distributions across both users and items, each user is allocated a budget B, representing the total number of interactions they can generate. These budgets are drawn from an exponential distribution, parameterised by a rate parameter β, which governs the degree of variation in user activity levels. This choice of distribution reflects the empirical observation that a small proportion of users tend to be highly active, while the majority exhibit relatively low levels of engagement.

$$B_i \sim \text{round}(\text{exponential}(\beta)) + c_b \tag{3}$$

In this formulation, c_b denotes a positive hyperparameter representing the expected number of interactions generated by each user. The user budgets are sampled from an exponential distribution, whose probability density function is defined as

$$f(x) = \beta^{-1}\exp\left(-x\beta^{-1}\right),$$

where β is the rate parameter, controlling the spread of user activity levels.

Having established user-specific budgets, the model defines a personalised probability distribution over items for each user. In practice, user behaviour is not random; empirical evidence from multiple domains indicates that users are more likely to engage with items they strongly prefer. Additionally, item popularity plays a significant role, with certain items receiving disproportionately high levels of interaction across the user population.

To capture these effects, the probability distribution over items for a given user i is constructed as a normalised element-wise product of two components. The first component is an item popularity vector, sampled from a power-law distribution to reflect the heavy-tailed nature of real-world item exposure. The second component is the latent preference vector $\tilde{R}_i$, derived from the probabilistic matrix factorisation stage, which encodes the user's intrinsic affinity towards each item. By combining these factors multiplicatively and applying normalisation, the model produces a distribution that jointly accounts for individual preferences and global popularity trends.

$$p_j \sim \text{Power}(a) + c \tag{4}$$

$$\mathbf{D}_i = \frac{\mathbf{p} \circ (\widetilde{\mathbf{R}}_i + \min_j(\widetilde{\mathbf{R}}_{ij}))}{\|\mathbf{p} \circ (\widetilde{\mathbf{R}}_i + \min_j(\widetilde{\mathbf{R}}_{ij}))\|_1} \tag{5}$$

In this formulation, p_j denotes the popularity weight associated with item j, while a represents the shape parameter of the power-law distribution governing item popularity. A positive constant c is introduced to ensure a non-zero baseline probability for all items, preventing the assignment of zero probability to any item. The operator $\circ$ denotes the Hadamard (element-wise) product between vectors or matrices of equal dimensionality, and $\|\cdot\|_1$ represents the ℓ_1-norm used for normalisation. The power-law distribution is defined as

$$f(x) = ax^{a-1},$$

capturing the heavy-tailed behaviour commonly observed in item popularity distributions.

For each user i, a total of B_i interactions are sampled without replacement from the user-specific distribution D_i. This process produces a binary masking matrix M, which determines the observed interactions within the dataset. The mask is subsequently applied to the latent preference matrix $\tilde{R}$ via element-wise multiplication, yielding the masked rating matrix

$$\tilde{R}^{(M)} = M \circ \tilde{R}.$$

In practical applications, user preference data is often discrete in nature, reflecting either implicit feedback, such as clicks or consumption behaviour, or explicit feedback in the form of ratings. Accordingly, the entries of $\tilde{R}^{(M)}$ are transformed to produce the final preference matrix R. This transformation may involve binarisation to represent implicit interactions, or discretisation and scaling to align with predefined rating schemes.

The overall data generation process is summarised in Algorithm 1. From a computational perspective, the dominant cost arises from the matrix factorisation stage, specifically the construction of the dense latent preference matrix $\tilde{R}$, which requires $\mathcal{O}(NM)$ operations. In contrast, the generation of the masking matrix M scales linearly with the number of observed interactions, and is therefore comparatively efficient.

Algorithm 1: Generating preference data with SPUP.

Data: N, M, β, c_B, a, c
for *each user* i *and item* j **do**
  Sample a rating $\widetilde{r}_{ij}$ *using Equation 2*;
end
for *each user* i **do**
  Generate a budget B_i *using Equation 3*;
end
for *each item* j **do**
  Generate a popularity $\mathbf{p}_j$ *using Equation 4*;
end
Set $\mathbf{M}$ to be an all-zero N by M matrix;
for *each user* i **do**
  for $k \leftarrow 0$ **to** B_i **do**
    Generate $\mathbf{D}_i$ using Equation 5;
    Sample an item: $j \sim \mathbf{D}_i$;
    Set $\mathbf{M}_{ij} = 1$;
  end
end
return $\mathbf{M} \circ \widetilde{R}$

It is important to note that, for large-scale datasets, the computation and storage of the full dense matrix $\tilde{R}$can become prohibitively expensive. In practical settings, such matrices may contain billions of real-valued entries, making in-memory processing infeasible. To address this limitation, the sparsification procedure is applied incrementally, either on individual users or on batches of users. The resulting sparse submatrices are then concatenated to construct the final dataset, significantly reducing memory requirements while preserving the statistical properties of the generated data.

4 Experimental Methods

To evaluate the effectiveness of the proposed Sparse Probabilistic User Preference (SPUP) model, we assess its ability to reproduce the key characteristics of real-world user preference datasets. For each dataset considered, the model's hyperparameters are adjusted to generate synthetic data that closely aligns with the observed statistical properties of the original data. These properties include structural and distributional attributes relevant to recommender system performance.

As a point of comparison, the clustering-based approach described in Section 2 is also implemented under the same evaluation framework. Synthetic datasets generated by both

methods are analysed and compared against the original datasets, enabling a systematic assessment of their relative fidelity.

In addition to statistical comparison, we further validate the utility of the SPUP model by training a standard collaborative filtering algorithm on the generated synthetic data. The results demonstrate that models trained on SPUP-generated data achieve substantially better performance than those trained on data produced using a naïve random masking strategy. This indicates that the proposed method not only captures the structural properties of real-world data but also preserves the underlying patterns necessary for effective learning.

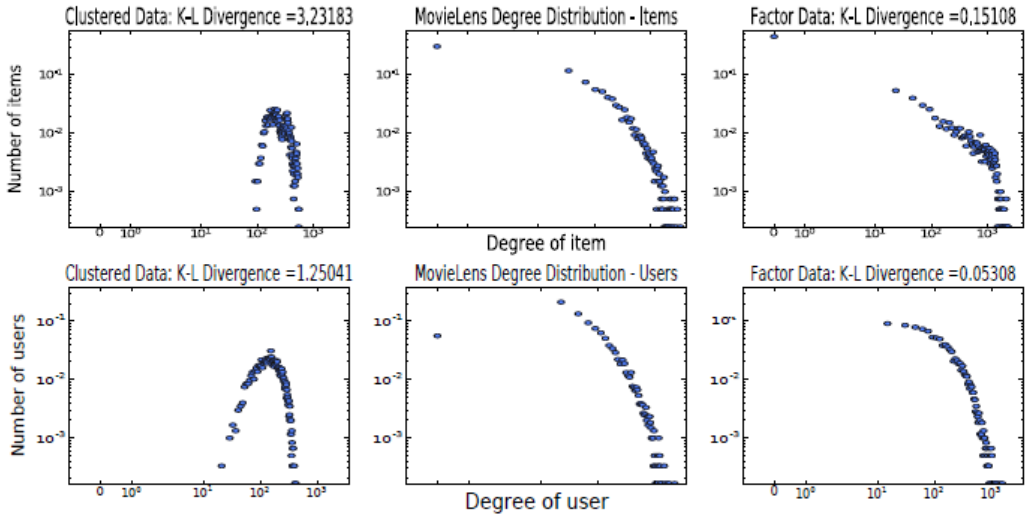

Figure 1illustrates the degree distribution observed in the MovieLens dataset, comparing results generated by the clustering-based method (left), the original dataset (centre), and the proposed SPUP model (right). The horizontal axis represents the degree, corresponding to the number of interactions associated with either items (top row) or users (bottom row), while the vertical axis indicates the frequency of items or users exhibiting a given degree. Each point therefore reflects the count of items or users corresponding to a specific degree value. Error bars are not displayed, as both methods exhibit negligible variance across repeated runs. Visual inspection reveals that the distributions produced by the SPUP model closely align with those of the real dataset. This observation is further supported by quantitative evaluation using Kullback–Leibler divergence, which indicates a smaller divergence between SPUP-generated data and the original dataset compared to that obtained using the clustering-based approach. These results suggest that the SPUP model more accurately captures the structural characteristics of real-world preference data.

To evaluate the proposed model, we consider multiple real-world datasets spanning diverse application domains, each exhibiting distinct user behaviour and consumption patterns.

MovieLens Dataset

The MovieLens dataset originates from an online movie recommendation platform where users provide explicit ratings on a discrete scale ranging from 1 to 5. These ratings are subsequently used to generate personalised recommendations. The version used in this study contains approximately one million ratings from around 6,000 users across 4,000 movies.

Million Song Dataset (MSD)

The Million Song Dataset captures user listening behaviour through implicit feedback in the form of play counts. The subset used here consists of approximately 1.4 million user–song interactions, covering 110,000 users and 160,000 songs. Play counts vary significantly, ranging from 1 to 923, reflecting heterogeneous listening behaviour.

Epinions Dataset

The Epinions dataset is derived from a product review platform where users rate purchased items on a scale from 1 to 5. It includes interactions from approximately 22,000 users and 300,000 items. In addition to ratings, the dataset incorporates social trust relationships between users, which influence the visibility and weighting of reviews.

Book-Crossing Dataset

This dataset originates from a book-sharing platform that tracks the movement of books across users through unique identifiers. It contains approximately 1.1 million ratings from 280,000 users on 270,000 books. Feedback includes both explicit ratings, on a scale from 1 to 10, and implicit interactions.

Comparison Measures

To assess the fidelity of the generated synthetic data, we compare it with real datasets across several structural and statistical attributes.

The first measure is matrix density, which reflects the proportion of observed interactions within the user–item matrix. Density serves as an indicator of the amount of information

available for modelling user preferences, with denser matrices generally providing more reliable estimates. Formally, for a matrix $R \in \mathbb{R}^{N \times M}$, density is defined as the ratio of non-zero entries to the total number of possible entries. When generating synthetic data, maintaining a density consistent with the original dataset is essential to avoid discrepancies in information content.

The second measure is the degree distribution, which characterises the connectivity structure of the data when interpreted as a bipartite graph. In this representation, users and items form two disjoint sets of nodes, and observed interactions correspond to edges. The degree distribution captures how many interactions each user or item has, thereby reflecting patterns such as user activity levels and item popularity. This is typically visualised using normalised histograms, where the horizontal axis represents the number of interactions and the vertical axis indicates the frequency of users or items exhibiting that level of activity.

The third measure is the Normalised Sorted Sum of Ratings (NSSR), which provides insight into how interactions are distributed across users or items. NSSR is computed as the proportion of total ratings attributable to each user or item and is analysed using a sorted distribution. This metric highlights concentration effects, revealing whether interactions are evenly distributed or dominated by a small subset of highly active users or popular items. Steeper distributions indicate stronger concentration, while flatter distributions suggest more uniform engagement.

In addition to these first-order statistical comparisons, we employ a second-order evaluation based on model performance. Specifically, a baseline probabilistic matrix factorisation model is trained on synthetic datasets and evaluated on held-out data. The results are then compared with performance obtained using real datasets, providing an indication of how well the synthetic data supports downstream learning tasks.

To quantitatively compare distributions, we utilise the Kullback–Leibler (KL) divergence, which measures the dissimilarity between two probability distributions. Given two discrete distributions, KL divergence captures the information loss incurred when one distribution is used to approximate another. To ensure numerical stability, Laplace smoothing is applied in cases where zero-probability values may arise.

Parameter Configuration

To generate synthetic datasets that closely resemble specific real-world datasets, the hyperparameters of the SPUP model are tuned through a structured search process. The procedure begins by adjusting parameters controlling sparsity, such as the rate parameter β and the baseline budget parameter, to match the density of the target dataset. Subsequently, user activity parameters are refined while maintaining the desired density level. Finally, item popularity parameters, governed by the power-law distribution parameters a and c, are calibrated to reproduce realistic popularity patterns.

In practice, the parameters β and c can be estimated directly from the data. The parameter cb is set to the minimum number of interactions observed across users, while β is derived from the average number of interactions per user, adjusted relative to cb. The selected hyperparameter values for each dataset are summarised in Table 1.

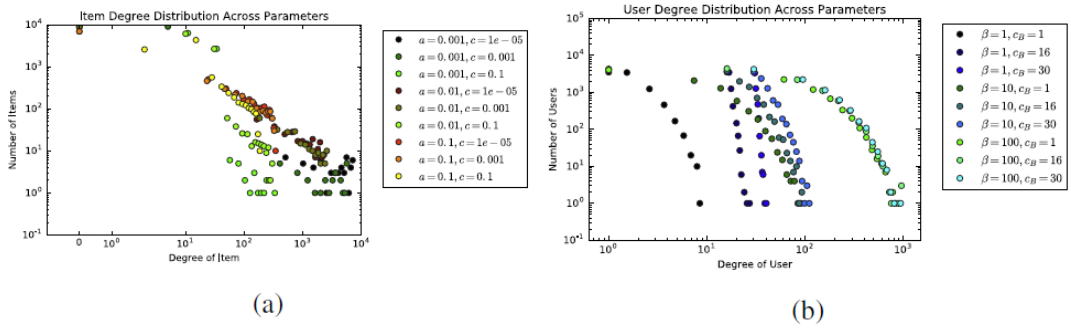

Figure 2 illustrates the sensitivity of degree distributions to variations in model hyperparameters. The left panel presents the degree distribution of items, while the right panel shows the corresponding distribution for users. These plots demonstrate how changes in hyperparameter values influence the structural properties of the generated data. Distinct patterns highlight the model's ability to control interaction behaviour across both users and items. For clarity, the figure is best interpreted in colour.

5 Results

Having outlined the methodological framework, we now present the empirical results of our evaluation. For consistency across methods, all datasets are converted to implicit feedback by binarising interactions, as the clustering-based approach does not support the generation of explicit rating values.

Dataset Comparison

Figure 1 presents the results obtained for the MovieLens dataset. The degree distribution

observed in the original data (centre panel) exhibits characteristics that lie between an exponential decay and a power-law distribution for both users and items. The clustering-based method fails to accurately reproduce these patterns, as evidenced by the mismatch in both user and item degree distributions (left panel). In contrast, the SPUP model closely aligns with the real data (right panel), achieving substantially lower Kullback–Leibler divergence values for both user and item distributions.

A similar trend is observed when analysing the Normalised Sorted Sum of Ratings (NSSR), as illustrated in Figure 4. The SPUP model consistently produces distributions that more closely resemble those of the real dataset, indicating its ability to capture both interaction structure and engagement concentration. Furthermore, Figure 2 demonstrates the sensitivity of the model to budget-related hyperparameters, showing that these parameters can be tuned effectively to replicate fine-grained distributional properties at both the user and item levels.

The Million Song Dataset presents a more challenging scenario due to its extreme sparsity, with a density on the order of 10^{-5}, and pronounced long-tail behaviour. While the average number of interactions per item remains relatively low, a small subset of items accumulates disproportionately high engagement. Despite these complexities, the SPUP model successfully captures the underlying structure of the data, as shown in Figure 5. The learned parameters remain interpretable and consistent with empirical observations, with the average user interacting with approximately ten items and item popularity following a plausible power-law decay. NSSR results for this dataset exhibit similar patterns to those observed for MovieLens and are therefore omitted for brevity.

Results for the Epinions dataset, presented in Figure 6, further demonstrate the robustness of the SPUP model. In this case, the model provides a significantly improved fit to the real data compared to the clustering-based approach, particularly in terms of user degree distribution. Performance for item distributions is comparable across both methods, though SPUP maintains a slight advantage in overall fidelity.

For the Book-Crossing dataset, both methods achieve reasonable approximations of the observed data. However, quantitative evaluation reveals that the SPUP model provides a more accurate representation. Specifically, the clustering-based method yields KL divergence values of 0.1457 for item distributions and 0.02235 for user distributions, whereas the SPUP model achieves substantially lower divergence for items (0.00565) while maintaining comparable performance for users. These results indicate that, even in cases where visual differences are minimal, the SPUP model offers a more precise statistical match to real-world data.

Parameter	Domain	Description	MovieLens	MSD	Epinions	Book Crossings
β	$\mathbb{R}^+$	Controls the distribution of user budgets	160	10.2	36.2	4
c_B	$\mathbb{N}^+$	Minimum number of ratings per user	15	3	2	0
a	$0 \leq a \leq 1$	Controls the distribution of item popularities	0.18	0.12	0.012	0.012
c	$\mathbb{R}^+$	Baseline probability of an item being rated	10^{-13}	0.02	0.02	0.02

Figure 3 Values of the hyperparameters selected for generating the four datasets considered in this study.

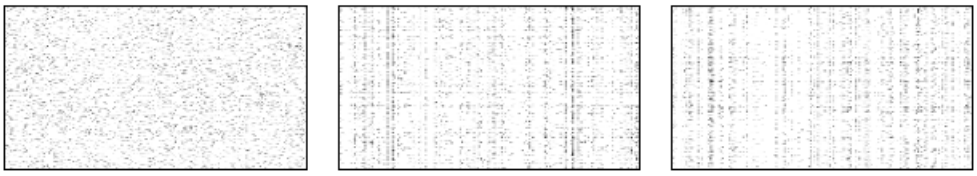

Figure 4 presents a visual comparison of observed user–item interactions in the MovieLens dataset, with users arranged along the rows and items along the columns. The left panel shows data generated using the clustering-based method, the centre panel corresponds to the original dataset, and the right panel displays the output of the SPUP model. To remove potential artefacts arising from ordering, both rows and columns were randomly permuted prior to visualisation. For optimal clarity, the figure is best viewed on screen.

Model Attributes

A key strength of the SPUP model lies in the interpretability of its parameters and their direct influence on the statistical properties of the generated data. Figure 2 illustrates how variations in key hyperparameters affect the degree distributions of both users and items.

Figure 2b demonstrates the impact of budget-related parameters on user degree distributions. In particular, these parameters govern the density of the generated ratings matrix by controlling user activity levels. Increasing the parameter β shifts the distribution towards higher values, effectively widening it and increasing the average number of interactions per user. Similarly, the parameter c_b, which represents the baseline number of interactions, produces a uniform shift in the distribution by adding a constant offset to user budgets. Together, these parameters provide intuitive and direct control over the sparsity and scale of the dataset.

The influence of hyperparameters on item degree distributions is more nuanced. While the overall position of the distribution is primarily determined by the user budget parameters, the shape is governed by the item popularity parameters a and c. As shown in Figure 2a, increasing the parameter a results in a redistribution of probability mass towards lower-degree values, effectively flattening the distribution in accordance with a power-law behaviour. This reflects a reduction in extreme popularity concentration. The parameter c introduces a baseline scaling effect prior to normalisation, which further moderates the

distribution. Specifically, increasing $c$leads to a horizontal shift towards lower degrees, reflecting a more uniform allocation of interactions across items. However, as αincreases, the influence of $c$diminishes, indicating that the shape of the distribution becomes increasingly dominated by the power-law parameter.

Beyond quantitative analysis, qualitative evaluation further highlights the strengths of the SPUP model. Figure 3 provides a visual comparison of the generated interaction matrices. The clustering-based method produces patterns that lack discernible structure, indicating limited variation in user behaviour and item popularity. In contrast, both the SPUP-generated data and the real dataset exhibit clear structural patterns, characterised by heterogeneous user activity and varying levels of item popularity. This visual similarity reinforces the ability of the SPUP model to capture realistic interaction dynamics.

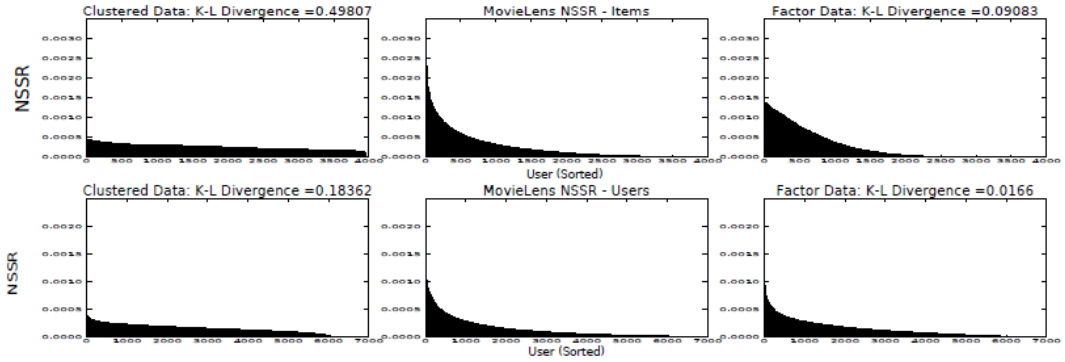

Figure 5 presents the Normalised Sorted Sum of Ratings (NSSR) for the MovieLens dataset, comparing the clustering-based method (left), the original dataset (centre), and the SPUP model (right). The top row illustrates the distribution across items, while the bottom row corresponds to users. The NSSR metric highlights how interactions are distributed, revealing patterns of concentration in user activity and item popularity. The SPUP-generated distributions more closely align with those of the real dataset, indicating a stronger ability to replicate underlying engagement patterns.

Learning with Synthetic Data

In the preceding analysis, we evaluated the fidelity of synthetic datasets by comparing their statistical properties with those of real-world data. We now extend this evaluation by examining the extent to which synthetic data supports downstream learning tasks. Specifically, we assess how effectively a collaborative filtering model trained on synthetic data can generalise, providing an indirect but practical measure of data realism.

The evaluation procedure is structured as follows. First, synthetic datasets are generated using the proposed model, or real datasets are used as a reference. Second, each dataset is partitioned into training, validation, and test subsets. Third, a standard collaborative filtering algorithm is trained and evaluated on these splits. In this study, probabilistic matrix

factorisation (PMF) is employed as the baseline model due to its widespread use and well-established performance in recommendation tasks.

Experiments are conducted using the MovieLens dataset, selected for its relatively higher density compared to the other datasets considered. We do not include comparisons with the clustering-based method in this stage of evaluation, as its lower fidelity in reproducing real-world statistical properties limits its effectiveness for meaningful learning-based assessment.

Results indicate that models trained on SPUP-generated synthetic data achieve substantially better performance than those trained on randomly generated data. In particular, evaluation using ranking-based metrics, such as Mean Normalised Discounted Cumulative Gain (NDCG), demonstrates that the synthetic data preserves meaningful interaction patterns that can be exploited by the learning algorithm. This suggests that the SPUP model not only captures structural properties of real datasets but also retains the underlying signal necessary for effective recommendation.

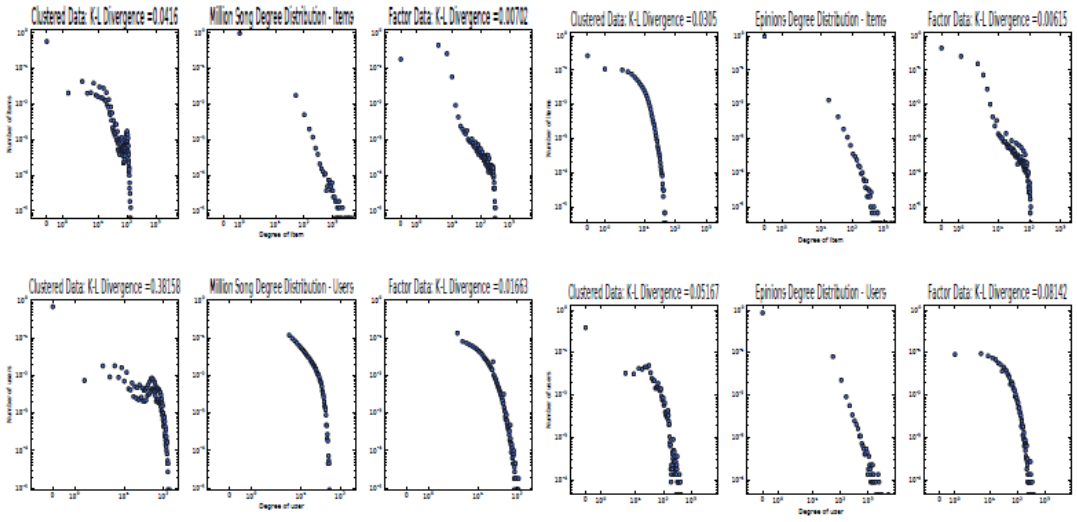

Figure 6 presents the degree distributions for the Million Song Dataset (MSD)

Specifically, the model trained on SPUP-generated data achieves a Mean Normalised Discounted Cumulative Gain (NDCG) score of 0.33 on the MovieLens test set, compared to 0.28 obtained using a baseline random predictor, where higher values indicate better performance. This improvement suggests that the synthetic data produced by the SPUP model retains meaningful structural patterns that can be effectively exploited by the PMF algorithm.

6 Discussion

This study introduced the Sparse Probabilistic User Preference (SPUP) model, a generative framework designed to produce realistic synthetic user–item interaction data. Empirical evaluation across multiple benchmark datasets demonstrates that the model is capable of reproducing key structural and statistical properties observed in real-world preference data. Furthermore, results from downstream learning experiments indicate that the generated datasets retain meaningful patterns that can be effectively exploited by standard recommendation algorithms, supporting the practical utility of the approach.

Comparative analysis between SPUP-generated data and real datasets highlights three core characteristics of the proposed model. First, the model exhibits a high degree of flexibility, enabling the generation of datasets with diverse structural attributes, including varying levels of sparsity, user activity, and item popularity distributions. Second, the model maintains strong interpretability, as the influence of individual hyperparameters on the resulting data can be understood in a direct and intuitive manner. This property is particularly valuable for controlled experimentation and sensitivity analysis. Third, the model demonstrates stability, consistently producing datasets with similar statistical properties under fixed parameter settings, and exhibiting low variability across repeated runs.

An important advantage of the SPUP framework lies in its modular design, which facilitates straightforward extensions. One potential direction involves the incorporation of auxiliary or side information, such as user or item features, into the data generation process. For instance, latent user representations derived from the matrix factorisation stage could be used to simulate social relationships between users, enabling the study of hybrid recommender systems that integrate both preference and network information.

Another promising extension is the integration of contextual factors into the generative process. By conditioning preferences on contextual variables such as time, location, or social environment, the model could be adapted to generate context-aware datasets. This would support the development and evaluation of more sophisticated recommendation systems that operate under dynamic and context-dependent conditions.

Overall, the proposed framework provides a foundation for generating rich, controllable synthetic datasets, enabling systematic experimentation in scenarios where real-world data is limited, restricted, or unavailable. By supporting both methodological development and practical evaluation, SPUP contributes to advancing research in recommender systems and related fields.

References

Aldous, D.J., 1981. Representations for partially exchangeable arrays of random variables. Journal of Multivariate Analysis, 11(4), pp.581–598.

Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B. and Lamere, P., 2011. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR).

Caron, F. and Fox, E.B., 2014. Sparse graphs using exchangeable random measures. arXiv preprint arXiv:1401.1137.

Cointet, J.P. and Roth, C., 2007. How realistic should knowledge diffusion models be? Journal of Artificial Societies and Social Simulation, 10(3).

Cassandra, T., n.d. POMDP File Repository. Available at: http://www.pomdp.org/examples/

Gopalan, P., Hofman, J.M. and Blei, D.M., 2015. Scalable recommendation with hierarchical poisson factorization. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).

Harper, F.M. and Konstan, J.A., 2015. The MovieLens datasets: History and context. ACM Transactions on Interactive Intelligent Systems, 5(4).

Hernandez-Lobato, J.M., Houlsby, N. and Ghahramani, Z., 2014. Stochastic inference for scalable probabilistic modeling of binary matrices. In Proceedings of the International Conference on Machine Learning (ICML).

Hoover, D.N., 1979. Relations on probability spaces and arrays of random variables. Technical Report, Institute for Advanced Study, Princeton, NJ.

Hu, Y., Koren, Y. and Volinsky, C., 2008. Collaborative filtering for implicit feedback datasets. In Proceedings of the IEEE International Conference on Data Mining (ICDM), pp.263–272.

ICAPS, n.d. International Planning Competition (IPC). Available at: http://www.icaps-conference.org/index.php/Main/Competitions

Koren, Y., Bell, R. and Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer, 42(8), pp.30–37.

Leskovec, J., 2008. Dynamics of large networks. PhD thesis. Carnegie Mellon University.

Newman, M.E.J., Strogatz, S.H. and Watts, D.J., 2001. Random graphs with arbitrary degree distributions and their applications. Physical Review E, 64(2), p.026118.

Pasinato, M., Mello, C.E., Aufaure, M.A. and Zimbrão, G., 2013. Generating synthetic data for context-aware recommender systems. In Proceedings of BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (CBIC).

RL-Glue, n.d. Reinforcement Learning Glue. Available at: http://glue.rl-community.org/

Rubin, D.B., 1993. Discussion: Statistical disclosure limitation. Journal of Official Statistics, 9(2).

Salakhutdinov, R. and Mnih, A., 2008. Probabilistic matrix factorization. In Advances in Neural Information Processing Systems (NeurIPS), pp.1257–1264.

Tang, J., Gao, H. and Liu, H., 2012a. eTrust: Discerning multi-faceted trust in a connected world. In Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM).

Tang, J., Gao, H., Liu, H. and Das Sarma, A., 2012b. eTrust: Understanding trust evolution in an online world. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.253–261.

Tso, K.H.L. and Schmidt-Thieme, L., 2006. Empirical analysis of attribute-aware recommender system algorithms using synthetic data. Journal of Computers, 1(4).

Ziegler, C.N., McNee, S.M., Konstan, J.A. and Lausen, G., 2005. Improving recommendation lists through topic diversification. In Proceedings of the International World Wide Web Conference (WWW).